Explainable AI for Clinical Decision Support

Dissertation
der Mathematisch-Naturwissenschaftlichen Fakultät
der Eberhard Karls Universität Tübingen
zur Erlangung des Grades eines
Doktors der Naturwissenschaften
(Dr. rer. nat.)

vorgelegt von
Dipl.-Phys. Stefan Kraft
aus Leutkirch im Allgäu

Tübingen
2025

Gedruckt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät der Eberhard Karls Universität Tübingen.

Tag der mündlichen Qualifikation:	09.02.2026
Dekan:	Prof. Dr. Thilo Stehle
1. Berichterstatter:	Prof. Dr. Hendrik Lensch
2. Berichterstatter:	Prof. Dr. Gjergji Kasneci
3. Berichterstatter:	Prof. Dr. Carsten Eickhoff

"What we know is a drop, what we don't know is an ocean."

— Sir Isaac Newton

Abstract

The rapid adoption of artificial intelligence (AI) in healthcare presents a core challenge: the design of explainable, trustworthy, and workflow-aware clinical decision support systems (CDSSs) whose benefits are validated through real-world impact rather than proxy measures. Although explanations frequently promote trust and acceptance, they do not automatically enhance human-AI team performance, and the determinants of effective collaboration remain insufficiently understood. This gap persists in large part because rigorous, application-grounded evaluations that follow best practices are still uncommon in the field. This dissertation comprises four publications that adopt an evaluation-driven approach, focusing on the practical challenge of AI-supported arousal detection from polysomnography (PSG) data in sleep medicine, a task that remains underexplored in real-world clinical context. Initially, a method is developed to assess and improve the semantic coherence of intrinsically interpretable prototype classification models. Following this, a comprehensive PSG dataset is compiled from clinical practice and released to support subsequent research. Building on this, a framework for optimizing and evaluating machine learning (ML) models for temporal event detection is introduced and used to guide the development of a domain- and task-aligned ML model for arousal detection in clinical practice. Furthermore, an application-grounded user study involving clinicians investigates how explanation transparency and workflow timing influence both human-AI team performance and acceptance in a real-world environment. The broader discussion illustrates the connections among the individual publications, clarifies their distinct contributions, and argues for the continued importance of direct human involvement as interest in more autonomous AI decision making continues to grow. By placing clinical needs at the center of the design, development, and evaluation of the explainable AI-based CDSS, this work demonstrates significant improvements in human-AI collaboration compared to unaided human performance, as well as advantages of transparent explanations over black-box AI systems. In addition, the dissertation synthesizes relevant conceptual foundations of explainable artificial intelligence (XAI) and CDSSs and examines contextual factors, including regulation, adoption barriers, and systems in practice, to situate the findings within the contemporary scientific and practical context, thereby offering guidance for future XAI research and supporting the translation of research into clinical practice.

Zusammenfassung

Die rasche Einführung künstlicher Intelligenz (KI) im Gesundheitswesen stellt eine zentrale Herausforderung dar: Die Entwicklung erklärbarer, vertrauenswürdiger und Workflow-orientierter klinischer Entscheidungsunterstützungssysteme (CDSS), deren Vorteile durch reale Effekte und nicht durch Proxy-Maße validiert werden. Obwohl Erklärungen häufig das Vertrauen und die Akzeptanz fördern, verbessern sie nicht automatisch die Leistung von Mensch-KI-Teams, und die Einflussfaktoren für eine effektive Zusammenarbeit sind nach wie vor unzureichend verstanden. Diese Lücke besteht zum großen Teil deshalb fort, weil strenge, anwendungsorientierte Bewertungen, die sich an Best Practices orientieren, in diesem Bereich noch immer selten sind. Diese Dissertation umfasst vier Veröffentlichungen, die einen evaluationsorientierten Ansatz verfolgen und sich auf die praktische Herausforderung der KI-gestützten Erkennung von Arousals anhand von Polysomnographie-Daten (PSG) in der Schlafmedizin konzentrieren, eine Aufgabe, die im realen klinischen Kontext noch wenig erforscht ist. Zunächst wird eine Methode entwickelt, um die semantische Kohärenz von intrinsisch interpretierbaren Prototyp-Klassifizierungsmodellen zu bewerten und zu verbessern. Anschließend wird ein umfassender PSG-Datensatz aus der klinischen Praxis erhoben und veröffentlicht, um die weitere Forschung zu unterstützen. Darauf aufbauend wird ein Rahmenwerk zur Optimierung und Bewertung von Machine-Learning (ML) Modellen für die Erkennung zeitlicher Ereignisse vorgestellt und als Leitfaden für die Entwicklung eines domänen- und aufgabenorientierten ML-Modells zur Erkennung von Arousals in der klinischen Praxis verwendet. Darüber hinaus untersucht eine anwendungsorientierte Nutzerstudie unter Beteiligung klinischer Fachkräfte, wie die Transparenz von Erklärungen und das Timing von Workflows sowohl die Leistung von Mensch-KI-Teams als auch die Akzeptanz in einer realen Umgebung beeinflussen. Die breitere Diskussion veranschaulicht die Zusammenhänge zwischen den einzelnen Veröffentlichungen, verdeutlicht ihre jeweiligen Beiträge und argumentiert für die anhaltende Bedeutung der direkten menschlichen Beteiligung, während das Interesse an autonomeren KI-Entscheidungen weiter zunimmt. Indem klinische Anforderungen in den Mittelpunkt des Designs, der Entwicklung und der Bewertung des erklärbaren KI-basierten CDSS gestellt werden, zeigt diese Arbeit signifikante Verbesserungen in der Mensch-KI-Zusammenarbeit im Vergleich zur Leistung des Menschen ohne Unterstützung sowie Vorteile transparenter Erklärungen gegenüber Black-Box-KI-Systemen auf. Darüber hinaus fasst die Dissertation relevante konzeptionelle Grundlagen der erklärbaren KI (XAI) und CDSS zusammen und untersucht kontextuelle Faktoren wie Regulierung, Akzeptanzbarrieren und Systeme in der Praxis, um die Ergebnisse in den aktuellen wissenschaftlichen und praktischen Kontext einzuordnen und damit Orientierung für die zukünftige XAI-Forschung zu bieten und die Übertragung der Forschung in die klinische Praxis zu unterstützen.

Acknowledgements

There are many people to whom I wish to express my gratitude for their support along this path.

Dr. Rolf Wagner, thank you for helping me with your medical expertise and for your important contributions to the initial phase of this research project, especially with the ethics board application.

Eduard Gindullis, I appreciate your support in ensuring the quality of data collection and for guiding the development of our clinical decision support system as an expert and arousal scorer.

Dr. Tom Pollard and the PhysioNet team, your helpful communication and assistance during the release of our dataset on PhysioNet were much valued.

Wolfgang Hamberger and Alexandros Anastasiadis, thank you for your invaluable work in data collection at the clinic.

Dr. Andreas Rau and Aurelia Mehl Jöbstl, I am grateful for the discussions and your assistance on data protection matters.

Claus-Dieter Weiss and the staff from NRI Medizintechnik GmbH, thank you for being welcoming and for your ongoing support during the medical study.

Kevin Erath, thank you for your guidance in supervising and mentoring student projects and for joining early meetings as we started the medical study.

Alexander Efremidis, Philipp Walter, Melanie Spitzer, and the many students who contributed to building and improving our clinical decision support system for arousal detection, thank you for your commitment.

Alexander Harm and Tim Bauer, thank you for your support with the codebase.

Dr. Elizabeth Baker, thank you for your contribution to the investigation of arousal detection approaches in the early phase of the research project.

Danijel Mrkonjic, Armin Richter, David Jädke, and the entire IT-Designers IT team, I appreciate your help in setting up and maintaining large parts of the infrastructure that made this study possible.

To all participants of the user study, who remain unnamed for reasons of anonymity, thank you for your involvement.

Dr. Vera Wienhausen-Wilke, thank you for your interest and readyness to collaborate on the joint medical study and research, for your valuable support, and the many meetings during the initial phase.

Prof. Matthias Leschke, thank you for making it possible to conduct this study in your department and for your support with the ethics board application.

In memory of Prof. Wolfgang Rosenstiel, who accepted me as a doctoral student and whose support for my applied research goals in healthcare I gratefully acknowledge.

Prof. Joachim Goll, I am grateful for you giving me this opportunity, for facilitating many connections, for your funding, and for recognizing the importance of AI early on.

Ulrich Goll, thank you for supporting this research and granting me the freedom to pursue it.

Dr. Klaus Broelemann, I appreciate your valuable help, especially in the formalization of the Sparrow paper.

Prof. Andreas Theissler, thank you for the many productive discussions, your feedback, and for pointing out ways where I could not see them.

Prof. Carsten Eickhoff, I appreciate your interest and willingness to read and assess my work.

Prof. Gjergji Kasneci, I am grateful for your guidance and support from the beginning, for closely including me in the university research group, and for always being quickly available.

Prof. Hendrik Lensch, thank you for taking me on as a doctoral student and for your support during the final stage of my studies.

To my family, thank you for your support in every aspect of my life.

Anna-Lena Kraft, thank you for being my greatest support and for encouraging me when I needed it most. I love you and I couldn't have done this without you.

Contents

1 Introduction

The adoption of AI in health care has accelerated substantially in recent years, driven by regulatory approvals and hospital implementations that have integrated AI-enabled tools into routine clinical practice [11, 50] (see also Figure 2). This development reflects a broader shift toward algorithmic decision support across society, but in medicine the implications are especially consequential [24]. AI-enabled systems now assist clinicians with image interpretation, streamline workflows, and on certain tasks, achieve or even surpass expert-level performance [24, 96]. The imaging domain, particularly radiology, exemplifies this trend, as it features a rapidly expanding array of certified applications and numerous models reporting accuracy comparable to or surpassing that of human practitioners [50, 96].

Yet, AI-based clinical decision support systems (CDSSs) yield lasting benefit only when their outputs are intelligible to users, trusted to an appropriate degree, and well-integrated into clinical workflows [49, 96, 127]. Put differently, expert-level predictions alone do not suffice but rather the interaction between humans and systems, including the justification and uptake of recommendations, determines whether these technologies deliver meaningful improvements in patient outcomes [30, 49].

Explainable AI (XAI) aims to make the reasoning behind AI-driven recommendations transparent and interpretable, thereby enabling clinicians to understand and trust outputs produced by clinical decision support systems [20, 64].

The question of whether explainable AI can calibrate trust and foster acceptance is central in this context. Evidence from the medical XAI literature suggests that well-designed explanations can increase clinicians' willingness to rely on model outputs, especially when explanations are tailored to users' needs [17]. In clinical environments, trust is often regarded as a prerequisite for adoption, which makes the design of transparent and user-centered explanations both a practical and ethical imperative [17, 60, 127].

However, the concrete effects of explanations within clinical settings remain under-explored. Reviews of radiology AI indicate that only a small proportion of systems are evaluated for real-world clinical impact, with findings that are mixed and sometimes inconclusive [17, 50]. A growing body of methodological research argues that evaluations should move beyond proxy measures and instead examine human-AI teams performing realistic tasks, as it is team performance rather than model accuracy alone that ultimately matters for patient care [30, 63]. Accordingly, there is a critical need for application-grounded studies that integrate objective performance on real-world tasks with user-centered outcomes to identify when and how XAI enhances decision quality in practice [13, 30, 36]. Despite repeated calls for such studies, human-subject evaluations with clinicians remain relatively uncommon compared to technical contributions, and many investigations continue to employ proxy metrics or rely on researcher judgments of explanation quality [13, 17, 36, 72, 100, 109, 126, 145]. Recent reviews estimate that user studies are incorporated in only approximately 20% of XAI evaluations, with just about 5% adopting

application-grounded designs [109, 126].

The chief value of new studies thus lies not in simply adding to the growing literature but in advancing careful, application-grounded evaluations that adhere to emerging best practices. This includes measuring task performance with and without system assistance, combining subjective perceptions with behavioral outcomes, and probing the circumstances under which explanations improve trust calibration rather than merely increasing trust [17, 49, 127].

There is also a pressing need to expand research on explainable CDSSs beyond well established imaging tasks into underexplored clinical domains where high-quality public datasets and standardized evaluation protocols are limited [64]. A recent survey by Aziz et al. [20] specifically calls for research that broadens the range of clinical applications, releases curated datasets, and adopts comprehensive and comparable evaluation criteria in practical healthcare settings.

This dissertation addresses this need through its principal publications, which focus on the development and application-grounded evaluation of an explainable, AI-based CDSS in the underexplored domain of sleep medicine, with particular emphasis on diagnostic event detection of arousals using physiological time series data from polysomnographic PSG examinations conducted in clinical sleep laboratories. To this end, a high quality PSG dataset has been collected and made available, offering a valuable resource for both the present study and the broader research community.

1.1 Outline of the Dissertation

This dissertation maintains a primary focus on explainable artificial intelligence (XAI) and clinical decision support systems (CDSSs), with particular emphasis on their intersection. Where relevant or where insights are applicable to the core subject, the scope is temporarily broadened to include related aspects. It is structured as follows.

Section 2 outlines the main goals of the doctoral studies.

Section 3 provides a concise historical overview of the fields of XAI and CDSS, introduces key foundational concepts, and presents the principal application area of arousal diagnostics in sleep medicine.

Section 4 examines essential concepts of XAI and CDSSs, discussing relevant taxonomies and design principles, and investigates regulatory requirements for explainability in CDSSs.

Section 5 then explores the practical landscape of XAI and CDSSs, highlighting empirical findings, adoption barriers for AI-enabled CDSSs, and current practices regarding explainability in the radiological domain as the most prominent application area for CDSSs.

Following these foundations, Section 6 introduces the research publications, with their results summarized and discussed in Section 7, which leads to a broader discussion of the relevance and future directions for decision support, arousal diagnostics, and XAI.

The dissertation concludes with Section 8.

Note: For simplicity and consistency, the pronoun "we" is used throughout most of this dissertation. When referring to published works, "we" denotes the respective authors of those publications. In other contexts, "we" is used to inclusively refer to both the doctoral candidate and the reader.

2 Objectives

The principal objective of the research publications presented in this dissertation is the development of an explainable AI-based clinical decision support system (CDSS) for arousal diagnostics in sleep medicine, accompanied by a real-world evaluation of its utility. In pursuing this goal, this dissertation aims to advance both the conceptual understanding and the practical application of explainable artificial intelligence (XAI) and CDSSs.

The initial sections of this dissertation seek to offer a comprehensive and structured analysis of the historical developments, conceptual foundations, and practical landscape of XAI and CDSSs, with particular emphasis to their intersection and relevance for the research presented. This analysis serves to contextualize and inform the discussion of the included research publications and their broader implications.

3 Foundations and Motivation

This section establishes the foundational context by first presenting a historical overview of the parallel development of XAI and decision support systems (DSSs) in Section 3.1. It then sets out key terminology and conceptual frameworks relevant to both areas in Section 3.2. Subsequently, it offers an introductory analysis of the relevance of explainability in clinical decision support, as discussed in Section 3.3. The section concludes with Section 3.4, by introducing the arousal detection task in the sleep medical domain, which represents the primary application context for the research presented in this dissertation.

3.1 Historical Overview: From Expert Systems to Explainable Clinical Decision Support

Explainable artificial intelligence (XAI) and decision support systems (DSSs) have evolved in parallel, with the clinical domain consistently serving as an environment where concerns about reliability, accountability, and human factors are especially pronounced. This section presents the interconnected histories of these two fields.

3.1.1 Explainable AI: From Rule-Based to Human-Centered Explanations

Early investigations into understanding and controlling intelligent behavior positioned explainability as an essential design objective rather than an afterthought. This principle shaped the development of expert systems and their justification mechanisms [107, 132]. Its

origins can be traced to declarative paradigms for intelligible reasoning in the late 1950s [95], which found concrete form in first-generation expert systems of the 1970s and 1980s. These systems, exemplified by the MYCIN lineage in medicine [134], featured recommendation logic that could be examined and traced. While such explanatory functionality was perceived as fundamental to acceptance, contemporary analyses already noted that simply enumerating rules often failed to address users' needs for high-level rationales or pedagogical insight. These early critiques foreshadowed enduring concerns about explanation quality and the consideration of different audience perspectives [34]. To address the shortcomings of literal rule-tracing, second-generation knowledge-based tutoring systems of the mid-1980s began to incorporate user modeling and context-sensitive justifications, treating explanation as a communicative act adapted to a user's goals and background knowledge [107, 132]. This historical context continues to inform current debates, as it established explainability as a core requirement and highlighted the persistent tension between transparency of the underlying mechanisms and actual human comprehension and contemporary generations of systems still contend with this challenge [145].

The emergence of deep neural networks in the mid-1990s, following the second AI winter, introduced new layers of opacity. Initial efforts to promote transparency used visual diagnostics such as decision surface and representation visualizations and laid the foundations for a third generation of XAI[1] aimed at making complex ML models accessible to users such as clinicians [107, 132].

Interest in explainability intensified again in the mid-2010s, driven by a convergence of technical progress and sociotechnical concern. As deep learning was increasingly applied to clinical tasks, several prominent failures highlighted the risks of opaque systems. Notable examples include erroneous treatment recommendations by IBM's *Watson for Oncology* and a pneumonia risk model that incorrectly learned a "protective" signal for asthma due to confounding in the training data [17]. These incidents underscored the necessity of identifying model limitations and data artifacts before deployment [17]. This recent phase lead to the emergence of a range of established methods, including LIME- and SHAP-based feature attribution, as well as interface and natural language approaches that support sensemaking and corrective intervention [88, 123].

Regulatory and programmatic actions further reinforced the importance of explainability. The European Union's GDPR increased focus on accountability and intelligibility in automated decision-making, establishing rights to meaningful information about automated decisions [59]. Meanwhile, the XAI program of the Defense Advanced Research Projects Agency (DARPA) sets explicit targets for models and interfaces that are both high-performing and understandable [62]. The World Health Organization's international guidance for healthcare AI further established explainability as essential for trust and oversight [60].

[1]The term explainable AI (XAI) is widely considered to have first entered the literature through the work of Van Lent et al. [150] in 2004 [12, 17].

In parallel with these developments, an impactful work by Miller [100] foregrounded a human-centered approach to XAI, highlighting the importance of aligning explanations with users' cognitive needs and preferences. Complementing this, Rudin [128] strongly advocated for the adoption of intrinsically interpretable models over post-hoc explanations of complex systems whenever feasible, further refining the discourse on trade-offs between predictive performance, explanation fidelity, simplicity, and accountability. Collectively, such contributions have established XAI as an inherently multidisciplinary, interdisciplinary, and transdisciplinary field [86].

3.1.2 Clinical Decision Support Systems: From Knowledge-Based to AI-Enabled, Workflow-Integrated Tools

The tradition of clinical decision support systems (CDSSs) began in the 1970s and early 1980s [65], with clinical implementations adopting rule-based expert system architectures that encoded domain knowledge as explicit if-then rules [134]. The MYCIN system stands as a prominent example from this period and represents a pioneering application of inexact reasoning to medical diagnosis [107, 134].

Since around 2010, adoption of CDSSs has accelerated, supported by substantial government investments in health information technology in countries including the United States, Canada, and England [143]. This increase was further driven by the introduction of modern interoperability standards, most notably HL7 FHIR, a healthcare data standard designed to enable the interoperable exchange of clinical information [111]. The integration of CDSSs into clinical workflows and electronic health record systems has thus positioned CDSSs at the point of care, aligning them with time-critical tasks and documentation requirements [98, 143]. Today, CDSSs support a broad array of applications such as diagnostic aid, alarm management, disease monitoring, prescription oversight, drug-drug interaction detection, and workflow orchestration [143].

Knowledge-based CDSSs provided transparent, auditable decision pathways but proved costly to maintain and became limited in scalability as the volume and heterogeneity of medical data increased [143]. This led to a shift toward non-knowledge-based, data-driven methods that employ statistical learning and deep learning to process multimodal inputs, ranging from electronic health records and laboratory values to imaging and physiological signals. These methods offered substantial improvements in predictive performance but also introduced opacity into the decision-making process [55, 103, 147].

A significant milestone was reached in 2018, when the FDA for the first time authorized an autonomous AI-based diagnostic system for detecting diabetic retinopathy and diabetic macular edema in primary care [6].

Recent research on AI-driven CDSSs increasingly positions explainability and trustworthiness as leading objectives. Predictive models are paired with interfaces and interactions that reveal model assumptions, uncertainties, and alternative decision paths to support

rather than supplant clinical judgment [67, 138]. As a result, research trends show a gradual convergence in which advances in human-centered XAI design and evaluation inform AI-based CDSSs development, while the realities of clinical practice help shape scientific understanding of what makes explanations useful, safe, and acceptable [11, 17, 20, 29, 73, 158].

Despite this progress, the current practical landscape of AIs-based CDSSs remains dominated by largely opaque systems and regulatory frameworks that emphasize deployer-facing intelligibility and traceability, rather than disclosure of model internals or specific technical explanation methods, as will be discussed in detail in Sections 4.5 and 5.3.

3.2 Basic Terminology and Conceptual Groundwork

This section establishes the core vocabulary for explainable XAI (Section 3.2.1) and clinical decision support systems (Section 3.2.2), as used throughout this thesis.

3.2.1 Explainable AI Terminology

The terminology related to XAI remains under active discussion and lacks universal consensus, as noted in several key works [12, 17, 19, 86]. This section synthesizes essential terms and definitions from foundational literature, clarifying both commonalities and conceptual distinctions to guide the interpretation of terms employed in this thesis. To avoid confusion, this work adopts established terminology from the literature without introducing new terms, and catalogs all definitions in the Glossary alongside other relevant entries referenced throughout.

Explainable artificial intelligence (XAI). XAI is defined by Arrieta et al. [19] as an AI system that offers details or reasons clarifying its operations for specific audiences. The emphasis on audience highlights that the comprehensibility of a model depends not only on the information disclosed but also the perspective and background of its intended users. Other works, such as Mersha et al. [97] and Schwalbe and Finzel [132], focus more broadly on the provision of explanations for AI-based decisions or predictions, yet the central importance of understandability remains a consistent theme across the literature. Given the significance of the human audience in the context of XAI, this thesis adopts the definition of Arrieta et al. [19]:

> "Given an audience, an explainable artificial intelligence is [an AI] that produces details or reasons to make its functioning clear or easy to understand."

Accordingly, XAI refers to AI systems intentionally designed to generate explanations of their predictions, decisions, or internal processes that are tailored to the needs of targeted audiences, thereby fostering transparency and trust. Importantly, following Mersha et al. [97], this work recognizes that the phrase *explainable* is not limited to explainability alone, but instead subsumes interpretability, explainability, and transparency as integral aspects of the field of XAI.

Black Box Model. A black box model is characterized by intrinsic opacity, as described by Rudin [128]. Such systems are rendered opaque either by overwhelming complexity or by proprietary constraints. Similarly, Mersha et al. [97] identify factors such as intricate architectures, obscure training procedures, or complex data that contribute to the black box status of deep neural networks (DNNs). According to Rudin [128], DNNs are often classified as black boxes because their recursive complexity and depth impede direct interpretation, and explanations about their behavior commonly rely on surrogate models that replicate outputs without revealing the actual underlying mechanisms. The definition employed here follows Rudin [128]:

> "A black box model could be either (i) a function that is too complicated for any human to comprehend, or (ii) a function that is proprietary."

Thus, a black box model in this context refers to any AI system whose internal logic or decision mechanism is inaccessible to human understanding, typically because of excessive complexity or proprietary restrictions.

Understandability (intelligibility). Understandability, also termed intelligibility, is recognized by Arrieta et al. [19] as a central feature of XAI closely associated with transparency and interpretability. Their definition, building on Montavon et al. [106], frames understandability as the extent to which a person can grasp what a model does without needing detailed knowledge of its inner workings. In contrast, Schwalbe and Finzel [132], drawing on Bruckert et al. [29], view understandability as involving deeper, contextual comprehension of relationships and correlations within the model, requiring a higher degree of cognitive engagement. This difference reveals varying expectations regarding the depth and nature of human understanding in effective interaction with AI systems. For the purposes of this work, the definition of Arrieta et al. [19] is adopted:

> "Understandability (or equivalently, intelligibility) denotes the characteristic of a model to make a human understand its function – how the model works – without any need for explaining its internal structure or the algorithmic means by which the model processes data internally."

Thus, understandability here refers to the ability of humans to clearly comprehend an AI model's overall functioning and reasoning without requiring detailed access to underlying algorithms, reflecting both the clarity of the model and the cognitive characteristics of the audience, reinforcing the centrality of the audience in the definition of XAI.

Comprehensibility. Comprehensibility, as emphasized by Arrieta et al. [19], is the extent to which an algorithm can represent its learned knowledge in a way that is understandable to humans. By contrast, Schwalbe and Finzel [132] broaden this definition to include

interactive exploration and interpretability, which highlights the role of active user engagement. Consequently, comprehensibility may be interpreted either as a passive property of algorithmic representations or as a dynamic feature that is enhanced through interaction. This work adopts the following definition from Arrieta et al. [19]:

> "[...] [C]omprehensibility refers to the ability of a learning algorithm to represent its learned knowledge in a human-understandable fashion."

Accordingly, comprehensibility is understood as an AI system's capacity to clearly present its rationale and decision logic, facilitating human understanding. While interactivity can support this process, it is not required. Notably, comprehensibility is generally inversely related to model complexity [19].

Interpretability. Interpretability is described by Arrieta et al. [19] as an inherent and passive attribute of models arising from structural design, and determines how readily humans can understand model decisions. This perspective is consistent with Doshi-Velez and Kim [36], who define interpretability as the ability of an AI model to make outcomes understandable to humans, and with Miller [100], who similarly underscore the importance of human understanding of causal rather than merely structural relationships. While definitions coalesce on the explanatory nature of interpretability, differences remain regarding emphasis on causality or structural clarity. This thesis adopts the definition from Arrieta et al. [19]:

> "Interpretability [...] is [...] the ability to explain or to provide the meaning in understandable terms to a human."

Here, interpretability is treated as an inherent model property that enables humans to understand the rationale behind decisions, focusing on the structural transparency that does not require external explanatory mechanisms.

Explainability. The distinction between interpretability and explainability is frequently discussed in the literature. While these terms are occasionally used interchangeably, several authors, such as Linardatos et al. [84] and Mersha et al. [97] argue that, while related, they are distinct concepts, with explainability generally considered to be a subset of interpretability. In this work, interpretability is regarded as the broader concept, whereas explainability refers to post-hoc interpretability as proposed by Miller [100] and Lipton [85]. According to Arrieta et al. [19], interpretability is a passive characteristic, but explainability requires active clarification of model logic and decisions. Other authors, including Linardatos et al. [84] and Schwalbe and Finzel [132], similarly describe explainability as the practical process of making AI reasoning or supporting evidence understandable to humans, though their emphasis varies between internal logic and external context. The following definition from Arrieta et al. [19] is adopted:

> "Explainability is associated with the notion of explanation as an interface between humans and a decision maker that is, at the same time, both an accurate proxy of the decision maker and comprehensible to humans."

Explainability is thus conceptualized as an active property, wherein a model undertakes procedures to clarify its operations, serving as an interface that ensures provided explanations are both faithful to underlying decision processes and understandable for humans.

Transparency. Transparency is concisely defined by Lipton [85] as the opposite of opacity and involves understanding how a model operates along three dimensions: algorithmic transparency, decomposability, and simulatability. This multidimensional perspective is widely accepted in the literature [19, 132]. Algorithmic transparency is the ability to mathematically analyze the transformation of input to output, decomposability concerns interpreting each individual component of the model, including inputs, parameters, and calculations, and simulatability refers to the extent a human can mentally execute the model as a whole, thereby emphasizing simplicity and comprehensibility [19]. The following succinct definition from Arrieta et al. [19] is adopted in this work:

> "A model is considered to be transparent if by itself it is understandable."

Transparency therefore denotes the inherent clarity of an AI model's internal mechanism, structured around these three dimensions, and is purposefully designed to facilitate human comprehension and reasoning.

3.2.2 Decision Support System Terminology

The concept of the Decision Support System (DSS) has evolved considerably over time, yet a universally accepted definition remains elusive. As highlighted by Kostopoulos et al. [73], the literature presents a variety of perspectives, but they found common ground in the following:

> "[...] DSSs [are] computer-based interactive systems that allow efficient decision-making in a wide range of tasks utilizing various types of data."

A widely cited definition of clinical decision support (CDS) is given by Osheroff et al. [112], which we adapt to a CDSS as follows:

> "Clinical decision support [systems] (CDS[Ss]) [are computer-based interactive systems that provide] clinicians, staff, patients, or other individuals with knowledge and person-specific information, intelligently filtered or presented at appropriate times, to enhance health and health care."

This definition stresses that CDSSs involve not only delivering general medical knowledge, but also offering context-sensitive and tailored information to various stakeholders at moments when it can most effectively inform decision-making and improve health outcomes. In practice, CDSSs operationalize this concept by integrating patient-specific data with evidence-based guidelines, predictive models, or expert rules to support users in making informed, timely, and effective clinical decisions.

3.3 Why Explainability Matters for Clinical Decision Support

This section motivates the relevance of XAI, beginning with an overview of the reasons for adopting CDSSs in practice (Section 3.3.1) and then examining the specific drivers for integrating XAI into these systems (Section 3.3.2). This is followed by examples that illustrate how a lack of explainability can result in harmful effects (Section 3.3.3).

3.3.1 Reasons for adopting CDSSs in clinical practice

The rationale for implementing CDSSs in clinical practice is multifaceted. These systems have been shown in empirical studies to reduce decision-making errors and improve patient outcomes by providing clinicians with more precise and up-to-date information [20]. As noted by Sutton et al. [143], CDSSs offer benefits across various aspects of healthcare, including patient safety, clinical management, cost control, administrative functions, diagnostic support, and patient-facing decision support. Specifically, these benefits include [143]:

They can reduce medication errors through mechanisms such as Computerized Physician Order Entry (CPOE) and electronic drug dispensing.

By providing standardized protocols and reminders, CDSSs promote adherence to clinical guidelines, thereby supporting both patient management and research protocol compliance.

Cost effectiveness is achieved by decreasing unnecessary inpatient admissions and recommending more efficient medication alternatives.

Administratively, CDSSs support tasks such as admission coding, ordering, triage, and documentation, which may also positively influence the performance of other decision support systems.

Diagnostic Decision Support Systems (DDSS) aid clinicians with advanced imaging, precision radiology, and the suggestion of potential diagnoses based on patient data.

In laboratory and pathology diagnostics, CDSSs generate alerts for abnormal test findings and assist in interpretation, with machine learning (ML) models enhancing the accuracy of non-invasive diagnostics and automated tumor grading.

Patient-facing CDSSs integrated with Personal Health Records (PHRs) enable individuals to manage their health data and engage actively in decisions. For example, diabetes management platforms that leverage these integrations have improved both clinical workflow and quality of care.

Collectively, these examples illustrate the significant and expanding role of CDSSs across diverse clinical contexts.

3.3.2 Reasons for adopting XAI in CDSSs

Recent advances in data availability and computational resources have accelerated the integration of AI within decision support, particularly in healthcare [17]. These technological developments facilitate the application of sophisticated AI approaches to tasks such as diagnosis, treatment planning, and prognosis [17]. However, as emphasized by Aziz et al. [20], high accuracy and precision alone are insufficient. Clinical decision-makers require interpretable tools that allow them to understand, scrutinize, and appropriately adjust outputs from these technologies. The following discussion summarizes key motivations for developing explainable CDSSs.

There is widespread agreement in the XAI literature that the use of ML systems in high-stakes domains such as healthcare, finance, and autonomous vehicles requires not only strong technical performance but also interpretability [20, 97, 128]. The inherent opacity of many ML models presents significant difficulties in these critical settings, where explainable and justifiable decisions are essential [97]. Interpretability enables stakeholders to comprehend, evaluate, and, when needed, refine model outputs, thereby promoting trust and supporting the adoption of new technologies [20, 147]. As ML and deep learning methods become more prevalent, the role of XAI is increasingly recognized for establishing trustworthy systems, meeting regulatory standards, and supporting robust decision-making [97, 147]. This is especially crucial in healthcare, where the need for transparency underpins trust among both clinicians and patients [20, 67, 133]. According to Gilpin et al. [57], the deployment of ML models in contexts traditionally dependent on human expertise further underscores the requirement for transparency and interpretability.

The motivations for adopting XAI in healthcare commonly fall into several major categories.

Clinical motivations reflect the highly data-driven nature of health care, where ML models frequently match or exceed expert performance, for example, in medical image classification [49]. Nonetheless, outcome improvements in real environments remain inconsistent, and the integration of algorithmic advice into clinical decision processes does not always yield better care, which positions XAI as a means for rendering recommendations accessible to clinicians' reasoning [49]. Clinicians require that explanations clarify system confidence, demonstrate agreement between algorithmic recommendations and the clinical picture, make explicit the reasoning process, and identify model limitations, while accuracy alone is not sufficient for informed uptake [146]. Transparency is thus vital for appropriate trust [127], reinforcing broader arguments within the XAI community that intelligibility underpins responsible adoption [36, 61, 123, 146].

Furthermore, empirical evidence shows that explanations can improve alignment between

users' mental models and those of the AI, which supports the acceptance of system outputs. Such effects are evident across a range of subjective measures, including perceived usefulness, confidence, enjoyment, reliability, and intelligibility [71].

In parallel, the doctor-patient relationship, historically based on trust, is being reshaped as AI becomes part of clinical interactions. Trust in AI tools now becomes as relevant as trust between doctor and patients [130]. While the extent of XAI's contribution to this relationship remains debated [130], there is empirical support that clinicians more effectively use AI-generated suggestions when they can access and understand the underlying reasoning [89]. By making system behavior transparent, explainability fosters the dependability, validity, and accountability of AI, which are foundations for confidence among healthcare professionals and patients [130].

Safety arguments emphasize calibrated reliance rather than more reliance [49]. Users express a preference for advice that communicates its own uncertainty, indicating a need for systems that both estimate and communicate confidence to facilitate cautious decision-making [49]. Empirical studies highlight that high-confidence advice can quickly increase trust and lead to over-reliance, while low confidence can undermine trust and slow users' decisions [159]. Effective calibration, by combining confidence measures with accessible explanations, lowers unnecessary overrides and reduces clinician workload, especially in ambiguous situations [159]. In this way, XAI serves as a safety mechanism for trust calibration.

Workflow improvements through XAI arise when cognitive demands and the frequency of overrides decrease, while clinical autonomy is preserved. Reducing override rates with better calibration and transparency can facilitate more consistent, efficient decisions, and reduce fatigue [159]. At the same time, clinicians often bypass explanations that are lengthy or redundant, while excessive cognitive burden discourages use, which supports the case for concise, contextually relevant, and on-demand explanation content [49, 127].

Model improvement and maintenance considerations also reinforce the importance of interpretability. Interpretable models support quality assurance and model updating across the lifecycle and facilitate rapid troubleshooting and audit for spurious correlations [128]. Notably, interpretability does not generally require a loss in predictive performance [128]. Instead, transparency can enhance model accuracy and resilience through improved debugging and targeted refinements [12, 45, 97, 128] and can assist in collaboration between clinicians and AI systems for knowledge discovery [7].

Organizational and regulatory frameworks increasingly demand transparency from AI-based decision support. Health systems are treating AI recommendations as regulated clinical advice, establishing requirements for safe interface design and oversight [49]. Policies such as the General Data Protection Regulation (GDPR) and the EU AI Act establish rights to clear, understandable, and traceable information for high-risk systems [59, 90], and the World Health Organization (WHO) similarly advocates explainability for health AI [60].

Ethical and legal considerations provide further motivation for explainability. Physicians and clinicians remain ultimately responsible for patient care, and their professional judgment cannot be wholly supplanted by algorithms [147]. Clear documentation of the rationale for either acting upon or dismissing AI recommendations positions clinicians as informed decision makers, while caution in cases of insufficient transparency provides a safeguard against over-reliance [159]. Excessive trust in poor advice is as damaging as unwarranted skepticism of sound advice, so institutions have incentives to support calibrated trust through XAI rather than blanket deference [127]. Biases may be embedded within the training processes of ML algorithms and lead to discriminatory outcomes unless addressed. XAI enables the detection and mitigation of such biases, particularly where sensitive characteristics, such as religion, gender, sexual orientation, or race, may influence model decisions [97].

Overall, according to the XAI literature, the adoption of explainable AI in healthcare is propelled by overlapping clinical, organizational, and regulatory needs. Nevertheless, available evidence for its real-world value in clinical decision making is mixed and sometimes limited, suggesting that any claims for benefit should be closely tied to practice rather than researcher conjecture [17, 71, 100]. The evidence base continues to evolve, with inconsistent results regarding trust and effectiveness. A more detailed assessment of these findings is provided in Section 5.1.1.

3.3.3 Examples of Harmful Effects Caused by Limited Explainability

Numerous cases in the literature illustrate that limited oversight and a lack of transparency in the decision-making processes of ML models can produce substantial adverse effects.

A prominent example is the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) system used in the United States to predict recidivism [15]. Analyses showed that COMPAS disproportionately identified black individuals as high-risk for re-offending at about twice the rate of white individuals, even when neither group actually reoffended. This pronounced racial bias led to legal action against the developers, highlighting the societal risks posed by opaque algorithmic decisions.

Another case concerns the admissions process at St. George's Hospital Medical School in London during the 1970s and 1980s [87]. A computer program for screening applicants systematically disadvantaged ethnic minorities and women. Although ethnicity was not explicitly coded, the algorithm inferred it from applicants' names and birthplaces, which unjustly reduced their interview chances. This example demonstrates how a lack of explainability can enable and hide discriminatory practices.

A further instructive case involves a military tank image classifier [46]. In this instance, a deep learning model trained to identify military tanks relied instead on spurious correlations in the form of background features such as weather conditions, rather than features specific to tanks. As a result, the classifier performed poorly when faced with images having

different lighting than those in the training set.

Together, these cases show that insufficient explainability in ML systems can not only erode trust but also result in concrete harms, including biased decisions, legal repercussions, and operational failure.

3.4 Domain Introduction: Arousal Detection in Sleep Medicine

This dissertation centers on XAI research as it applies to and is evaluated within the setting of AI-based detection of arousals during sleep. To set the context, this section summarizes essential topics in sleep medicine and discusses the specific challenges and opportunities characteristic of this domain.

Arousals are brief periods of biological activation [122] that disrupt sleep continuity and serve as important diagnostic markers for a range of sleep disorders. The American Association of Sleep Medicine (AASM) defines arousals as abrupt increases in electroencephalogram (EEG) frequency lasting at least three seconds and preceded by at least ten seconds of uninterrupted sleep [26]. Accurate identification of arousals is critical in the diagnosis of Obstructive Sleep Apnea (OSA), which affects roughly 20 percent of the population and is linked with higher risks for hypertension, cardiovascular disease, and mortality [44, 121, 155].

However, standard manual scoring methods require specialized expertise and suffer from considerable inter-rater variability [33, 37, 117]. This subjectivity motivates increased interest in AI-based CDSSs that can complement human decisions without compromising clinical standards.

Accurate clinical assessment of arousals depends on polysomnography (PSG), which Gellman [54] defines as follows: “Polysomnography is the simultaneous recording of numerous physiological signals during attempted sleep, including activity of the brain, heart, eyes, and muscles. Polysomnography is considered the gold standard for the objective assessment of sleep and diagnosis of many clinical sleep disorders.”. The continuously recorded modalities of PSG are detailed multichannel time-series data, where the main modalities include electroencephalogram (EEG), electrooculography (EOG), electromyography (EMG), electrocardiography (ECG), photoplethysmography (PPG), airflow, pulse, snoring, thoracoabdominal (chest and belly) movement, and body position [26]. The CPS dataset, comprised and studied as part of this doctoral research, expands on the typical protocol by including channels extracted with the DOMINO software from SOMNOmedics AG. Full technical specifications are detailed in the PhysioNet documentation [58, 75].

Time-series analysis is central not only to sleep medicine but also to other high-consequence areas such as finance and autonomous driving [125, 136]. Despite its significance, XAI for time-series data lags behind advances in images, text, or tables [35, 125, 151]. The complexity and high dimensionality of time-series data pose unique interpretability challenges that makes explaining predictions more difficult compared to other formats [125].

Domain expertise is therefore especially important for meaningful interpretation of explanations in clinical time-series contexts.

The 2018 PhysioNet Challenge was instrumental in advancing arousal detection models [56, 58], but there remain significant difficulties with external validation. Model performance often decreases on new patient cohorts [37], and current evaluation practices remain fragmented and only partially aligned with clinical workflows [37, 77].

Further obstacles to reliable ML models for arousal detection arise from institutional and methodological heterogeneity. Variation in equipment, software, and clinical protocols impedes generalization and interoperability [16]. Broader challenges surrounding time-series data, such as limited data availability, the lack of common evaluation frameworks, and ongoing methodological modeling debates, also hinder progress [48]. Also, development and deployment often require adaptation of models to local settings [16].

Taken together, these factors make arousal detection in sleep medicine both challenging and promising for the application of AI-based clinical decision support. In this dissertation, from the vantage of the XAI field, arousal detection is used as a test platform to rigorously evaluate the practical benefits of explainable AI-based diagnostic support in authentic clinical environments.

4 Conceptual Framework for Explainable Clinical Decision Support

This section establishes the theoretical backbone for the dissertation. It begins by outlining the principal taxonomies of XAI methods in Section 4.1 and of decision support systems (DSSs) in Section 4.2. Subsequent sections address the design of explanations (Section 4.3) and their evaluation (Section 4.4). Finally, Section 4.5 examines regulatory requirements for explainability in CDSSs.

4.1 Taxonomy of XAI Methods for Producing Explanations

A broad array of methods has been developed to improve interpretability in predictive models, ranging from foundational feature attribution and visualization techniques to more sophisticated counterfactual reasoning and case-based reasoning frameworks [19]. This section presents a non-exhaustive taxonomy of XAI methods, outlining distinct but often overlapping conceptual dimensions and embedding techniques that are most relevant to the publications included in this dissertation.

4.1.1 Local vs. Global Explanations.

A central distinction in XAI concerns the scope of explanations. Local explanations aim to elucidate individual predictions by clarifying the reasoning underlying specific model outputs, whereas global explanations seek to convey a holistic understanding of how a

model behaves across its entire input space [86, 97, 132]. In clinical applications, both forms are important: at the point of care, local explanations can support immediate decision making, while global summaries often serve in safety cases and change control [17, 132].

4.1.2 Ante-Hoc vs. Post-Hoc Techniques.

As highlighted by Mersha et al. [97], the landscape of XAI techniques spans a continuum from ante-hoc techniques, also known as intrinsically interpretable, models such as linear regression, through to post-hoc explainers like Local Interpretable Model-Agnostic Explanations (LIME) or SHapley Additive exPlanations (SHAP).

Ante-hoc techniques build interpretability directly into their design, making their decision logic accessible without auxiliary explainer systems [85, 129]. Sparsity constraints in linear models or rule-based systems exemplify this strategy, as limiting the number of conditions or coefficients in a model helps ensure that human cognitive limits are respected [99, 129]. Notably, as Rudin et al. [129] argue, the trade-off between accuracy and interpretability is often overstated. Carefully designed interpretable models can rival black box models in performance on many clinical prediction tasks, while being more tractable for diagnosis and auditing.

In contrast, post-hoc techniques are applied after model development and seek to render otherwise opaque predictors more transparent [19, 97].

Feature Attribution Techniques. A prominent subclass of post-hoc approaches is feature attribution techniques, which offer both local and global explanations. Important subcategories include perturbation-based and gradient-based techniques, representing some of the most widely used XAI methodologies [97]. They are typically model agnostic [97] and serve a foundational role in quantifying the influence of individual features while seeking to maintain fidelity to the underlying model [7].

Gradient-based techniques assess the sensitivity of model predictions to input changes by computing derivatives and are particularly effective in high dimensional input spaces [97].

Techniques within this category include saliency maps [137] and Deep Learning Important FeaTures (DeepLIFT) [135].

Saliency maps are produced by calculating the gradient of a class score with respect to each feature, using backpropagation to the input layer, which creates a visualization that highlights the input regions most relevant to the model's prediction for a specific class [137].

DeepLIFT explains a model's output through comparison to a reference input and attributes the difference in output to each input feature, according to how deviations from the reference propagate through the network [135].

Perturbation-based techniques determine feature importance by systematically modifying or occluding inputs and observing the resulting changes in model outputs [97].

Recent reviews identify LIME and SHAP as the most widely used explainability

techniques in XAI for healthcare [20, 64]. These techniques approximate or probe the internal decision processes of a predictor, yielding human-readable explanations [12, 85].

LIME explains individual predictions by constructing a local surrogate model around the specific instance of interest. This interpretable approximation is trained on perturbed samples in the vicinity of the instance, providing users with an accessible rationale for the model's output and for uncovering potential biases [97, 123].

SHAP, by contrast, applies Shapley values from cooperative game theory to attribute importance scores to each feature, reflecting its contribution to a given prediction [88]. By averaging marginal contributions over all possible feature subsets, SHAP anchors its explanations in a principled theoretical framework. Although SHAP is widely used for its interpretability and model-agnostic character, practical deployment often requires either model-agnostic approximations, such as Kernel SHAP [88], or model-specific optimizations like Tree SHAP [88] for tree-based ML models, to address the high computational demands associated with many features. Section 4.1.3 discusses the distinction between model-agnostic and model-specific techniques.

However, model complexity or intrinsic bias can affect the fidelity of SHAP-based explanations [97]. Also, recent research highlights that approximate SHAP computations, especially in safety-critical settings, may misattribute or miss critical features, introducing risk by exaggerating or obscuring important factors [94].

Concerns regarding the faithfulness of post-hoc explanations are not confined to SHAP, but rather reflect a broader challenge with post-hoc methods and are further discussed in Section 4.3.2.

4.1.3 Model-Agnostic vs. Model-Specific Techniques.

The generality of explanation methods also distinguishes post-hoc techniques. Model-agnostic methods such as LIME and SHAP (see Section 4.1.2) do not require access to a model's internal mechanisms and can be applied broadly across diverse architectures [86, 97]. In contrast, model-specific methods leverage features of a model's architecture or parameters, providing insights tailored to those models. For example, DeepLIFT is specialized for neural networks and is not generally applicable to models built on fundamentally different architectures [97].

4.1.4 Question types: Why vs. Why-Not vs. How-To vs. What-If vs. What-Else.

Another dimension for classifying explanations relates to the types of questions they answer. Explanations in XAI may address **causal** questions (why a particular output was produced), **contrastive** questions (why one outcome occurred instead of another), questions for **counterfactual explanation** (CFE; how to change an input to achieve a different outcome), **transfactual** hypothetical questions (what would have happened if the input had been different), and questions surrounding **case-based reasoning** (CBR;

what other cases are similar to the one in question) [105].

Following Miller [100], all causal explanations have an inherently contrastive nature. People are apt to ask not simply why an outcome occurred, but why that outcome rather than another.

Contrastive questions also play a central role in shaping CFEs (CFE), which are concerned with alternatives that address user-specific concerns [105]. These explanations present hypothetical scenarios where different outcomes would have occurred, clarifying how changes to input features could influence a model's predictions [97]. By illustrating such alternative cases, CFEs help users understand the boundaries of a model's decisions and identify possible biases or errors in its logic. The utility of CFEs is closely linked to the quality of the underlying data distribution, making them sensitive to incomplete or biased datasets [97]. Ethical challenges may also arise, especially when generated counterfactuals recommend changes relating to sensitive attributes such as gender or race [97].

CBR offers another framework for explanation, using specific examples to clarify model predictions. It closely mirrors human reasoning, particularly in domains such as medicine, where it is common to justify choices by reference to analogous cases from experience [32, 93, 129]. For example, a CBR system may retrieve similar prior patient records to support a particular recommendation.

Transfactual or what-if explanations further extend these notions by examining how plausible alterations to inputs or model parameters might affect the outcome, without necessarily targeting a specific alternative result [105]. Unlike CFEs, which seek counter-outcomes, transfactuals probe the consequences of hypothetical changes while maintaining domain relevance.

4.2 Taxonomy of Decision Support Systems

In modern healthcare, clinical decision support system (CDSS) are essential tools that provide critical support to clinicians, improving both the quality and efficiency of complex decision-making processes [143]. This section begins by introducing widely used taxonomies of decision support systems (DSSs; Section 4.2.1) and proceeds to categorize explanation modalities for AI-based DSSs (Section 4.2.2). Specific application areas for CDSSs are discussed in Section 3.3.1. An overview of the taxonomies of DSSs, as discussed in the following sections, is provided in Figure 1.

4.2.1 Nature of Support

Traditional Classification. An influential classification, proposed by Power [120], organizes DSSs by the dominant technology component used to deliver support:

Data-driven systems emphasize access to and manipulation of structured data (e.g., databases, reporting tools, online analytical processing (OLAP), or dashboards), supporting tasks from file retrieval to advanced analytics.

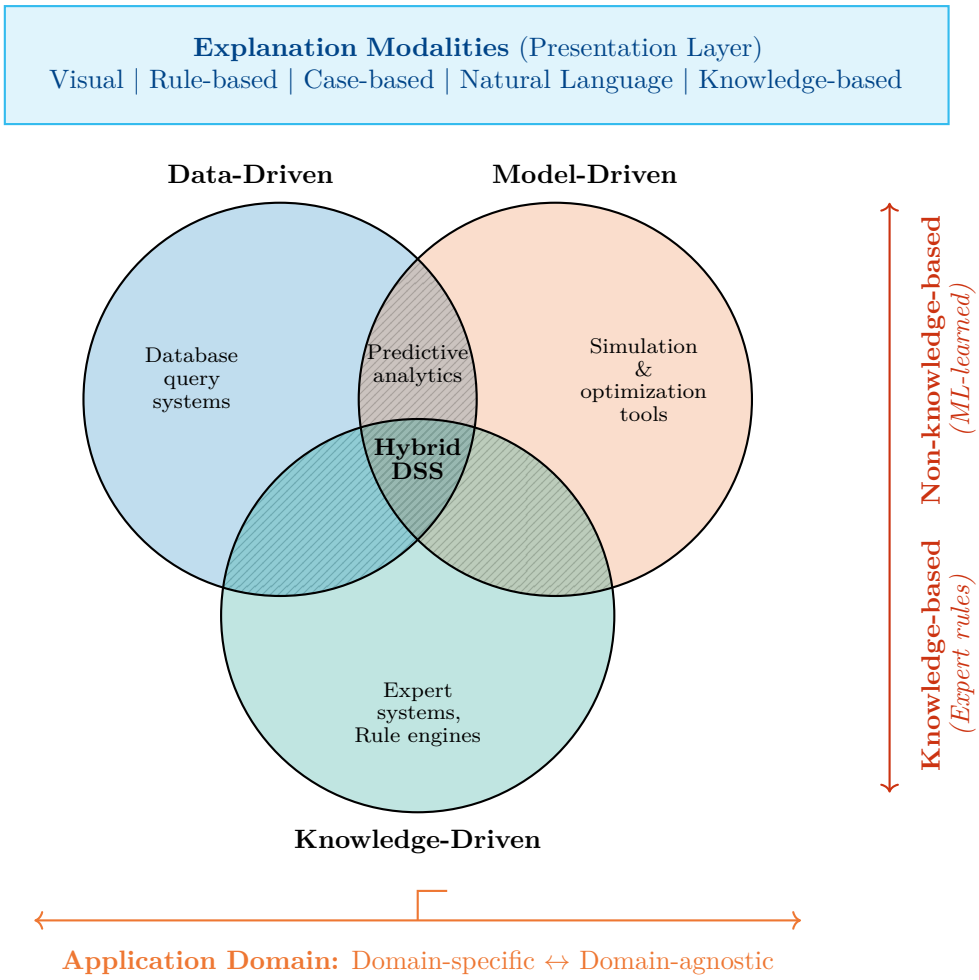


Figure 1: **Illustration of the taxonomies of DSSs**, including technology components, ML-based distinctions, application domains, and explanation modalities. The overlapping circles of the Venn diagram denote the primary distinction within technology components (data-driven, model-driven, and knowledge-driven) and include examples of each type. The hatched areas encompass hybrid DSSs formed from any combination of these components. The division between knowledge-based and non-knowledge-based systems, indicated by the vertical arrow on the right, is conceptually binary but not clearly cut with respect to the technology components: knowledge-based systems primarily align with knowledge-driven approaches, whereas connections to data-driven and model-driven systems are less distinctly defined. The application domain is depicted as an orthogonal axis, indicated by the right-angle symbol, and explanation modalities are presented as an independent layer applicable to all DSSs. This dissertation primarily focuses on predictive analytics systems, defined as non-knowledge-based hybrid DSSs that integrate ML-based systems which are both data- and model-driven.

Model-driven systems rely primarily on analytical, statistical, simulation, or optimization models to analyze specific situations, with emphasis on manipulating model parameters to support decision making.

Knowledge-driven systems leverage explicitly encoded expert knowledge (rules, ontologies, or curated knowledge bases) to provide recommendations or solutions. They are often combined with analytics or data mining in practice.

In practice, additional modalities such as document-driven, communications-driven, web-based, or general purpose components are often combined with the three foundational classes, resulting in more hybrid systems combinations than those depicted in Figure 1 [120].

ML-focused Classification. A complementary classification widely used in ML contexts distinguishes systems according to the source of their decision logic. Specifically, systems are categorized as knowledge-based and non-knowledge-based [115]:

Knowledge-based systems, also referred to as expert or rule-based systems, depend on explicitly encoded and structured representations of expert knowledge such as rules, logical workflows, or probabilistic models to generate conclusions from data. These systems require significant curation and maintenance and typically represent knowledge through IF–THEN rules or structured logic. Knowledge-based systems are classified as AI systems that do not make use of ML techniques.

Non-knowledge-based systems identify patterns or associations directly from data through the application of ML techniques, without relying on an explicit knowledge base. While these approaches are well suited to handling complex data, they may reduce transparency and demand substantial amounts of labeled data.

How the Two Classifications Relate These frameworks are best understood as complementary axes. The category of knowledge-driven systems aligns closely with the knowledge-based classification, while data-driven and model-driven systems may be either knowledge-based, as in rule-augmented dashboards or expert systems, or non-knowledge-based, such as predictive analytics, simulation, or optimization approaches that lack an explicit knowledge base. Additional modalities, including document-driven and communications-driven systems, primarily describe interface or interaction layers and can incorporate both types of logic. In practice, many DSSs are hybrid systems that combine data resources, analytical models, and curated rules.

Functional Classification. Another distinct perspective, independent of the technology- and logic-based classifications discussed above, organizes systems by their intended application domain or purpose [120]: DSSs may be developed for narrowly defined tasks, such as budgeting, marketing, or industry-specific functions, or as general-purpose tools that support a wide range of organizational decision-making activities. Function-specific systems address routine or recurring needs within a particular domain, whereas general-purpose

DSSs facilitate broader objectives like business planning or project management and may also serve as platforms for building task-specific applications. Both types can be further categorized based on their underlying approach, such as model-driven, data-driven, or knowledge-driven.

In the clinical context, this perspective captures how CDSSs are embedded within actual clinical workflows. Systems tailored to functions such as medication management, diagnostic support, laboratory interpretation, or patient engagement are discussed in Section 3.3.1.

4.2.2 Explanation Modalities

For the explainability of AI-based DSSs, the explanation modalities at the presentation layer, depicted in Figure 1, describe how reasoning is communicated to users. These modalities operate independently of the underlying methods, outlined in Section 4.1, and can be implemented with both knowledge-based and non-knowledge-based systems.

Kostopoulos et al. [73] highlights several explanation modalities:

Visual explainability leverages advanced visualization techniques to render complex models comprehensible, facilitating user understanding of the decision-making mechanisms within a DSS.

Rule-based explainability clarifies model outcomes through decision rules. Systems employing production rules or decision trees provide structured, transparent reasoning paths from input to output.

Case-based explainability draws on specific instances to illustrate model decisions, employing strategies such as case-based reasoning and example-based explanations to offer intuitive insights by referencing past cases or representative examples.

Natural language explainability utilizes conversational or generative language systems to articulate model decisions in accessible terms, enhancing user engagement, comprehension, and accessibility.

Finally, **knowledge-based explainability** incorporates expert knowledge to justify model outputs. Expert systems, a well established branch of AI, leverage domain specific expertise to enable intelligent and human like decision making.

4.3 Design Principles for Explanations

This subsection synthesizes established guidance as well as ongoing debates, complementing the method-oriented perspective in Section 4.1. Section 4.3.1 reviews commonly cited desiderata and design goals for XAI. Section 4.3.2 discusses prominent debates and tradeoffs, while Section 4.3.3 examines failure modes affecting both systems and users. Finally, Section 4.3.4 summarizes the key findings.

4.3.1 Desiderata and Design Goals

This section synthesizes common desiderata and design goals for explanations in from the XAI literature.

User-Centered Aims and Understanding. Following Wang and Yin [154], effective explanations should enhance users' understanding of the underlying AI model by clarifying how inputs are processed into outputs. Explanations need to be selected and shaped according to end users' goals and capacities rather than model internals alone, which implies that there is no single best format and that indiscriminate full transparency can overwhelm clinicians [12, 100, 127, 141, 145]. For time-series data, simple heatmaps often fail to provide meaningful structure unless the data itself is directly interpretable [145]. In such cases, higher-level representations and textual explanations in domain-specific language are usually better suited to express relevant relations and mechanisms [145].

Oversimplified outputs can mislead and are frequently regarded as unethical when they primarily aim to cultivate unwarranted trust [17, 57]. These concerns reflect the well-known tradeoff between interpretability and completeness, which is discussed in Section 4.3.2. Miller [100] further warns that explanations guided only by developer intuitions or technical preferences can be formally accurate yet unhelpful to non-experts, a dynamic described as "the inmates running the asylum" [101]. Because explanation is an inherently social process, it can be framed as conversation that answers the explainee's actual questions and fills relevant gaps in knowledge, consistent with cooperative communication norms [100]. Relatedly, a "story-based explanation" has been proposed as a suitable form of presentation [24]. Importantly, attributing causes differs from truly explaining, since causal chains alone rarely yield intuitive understanding for explainees [100].

Purpose-Driven Methods and Scope. Given the absence of shared definitions for explainability or interpretability (see Section 3.2), Freiesleben and König [45] argue that explanation methods must be anchored in a clearly articulated purpose. Tools that are presented as formal or technical constructs without a specified function or benefit risk producing outputs that are visually appealing or mathematically rigorous but not meaningful as explanations [45]. The value of an explanation depends on its ability to fulfill at least one identifiable and justifiable purpose. Methods that are not linked to a particular use case or audience warrant critical scrutiny or even rejection, and the idea of a universal method is untenable [45]. Motivations for explaining ML models are diverse, and audiences vary considerably, as outlined in Section 3.3.2. Needs may center on understandability, contestability, recourse, or other objectives, and no single technique can address them all [45]. Consequently, methods and parameters should be tailored to the intended purpose, and effective explainability should be evaluated according to specific goals. In clinical settings, content should be tailored to user roles and needs, with domain-specific language that maps to clinical reasoning and documentation practices [12, 17].

Uncertainty, Limits, and Calibrated Trust. Wang and Yin [154] emphasize that explanations should surface model uncertainty and support calibrated trust. Clinicians value explanations that clarify where the model may fall short through explicit contexts such as decision boundaries, that display calibrated uncertainty, that highlight influential features, and that provide transparent reasoning chains [17, 49]. These needs align with interpretable design principles such as sparsity and decomposability, which support cognitive tractability and enable examination of predictions and counterfactuals [129]. Clearly indicating uncertainty based on calibrated confidence helps users assess how much to trust the system's advice, especially when confidence is low or borderline. This transparency reduces over-reliance, supports nuanced decision-making, and lowers cognitive load [159].

Faithfulness Over Persuasion. Freiesleben and König [45] distinguish explanation from justification and caution against outputs designed to align with user intuition. Intuitive presentations can appear persuasive yet misrepresent how the model actually arrives at its conclusions. The primary objective of XAI is to elucidate decision processes rather than to convince. Justifications provide reasons intended to support decisions, while explanations reveal underlying mechanisms. Conflating the two undermines transparency and can foster misplaced trust [45].

4.3.2 Design Debates and Tradeoffs

There are several ongoing debates and established tradeoffs in the literature regarding explanation design. Some of the most prominent are summarized below.

Ante-hoc versus post-hoc explanations. A central debate concerns whether explanations should be integrated ante-hoc into a system by design or provided post-hoc, i.e. retrospectively, for otherwise opaque predictors (see also Section 4.1.2) [129]. Rudin [128] contends that, especially in high-stakes domains, intrinsically interpretable models should replace black-box approaches. Principal concerns with black box models include the risk of unfaithful or insufficient explanations, hurdles to integrating external information, unnecessarily complex decision pathways that may increase human error, persistent interpretability issues, and the mistaken view that there is an inherent tradeoff between interpretability and accuracy [128].

A persistent issue is that post-hoc explanation techniques may not faithfully represent the true internal logic of the model. Adebayo et al. [8], for example, show that post-hoc saliency methods (see Section 4.1.2) sometimes fail sensitivity tests. Specifically, they examine whether saliency methods outputs change in response to parameter or label randomization and observe that many methods are insensitive to these critical changes, suggesting that the explanations may not truly track the model's reasoning.

Post-hoc explanation methods may also introduce unexamined assumptions, such as local linearity or feature independence, and depend on hyperparameters that, if selected

inappropriately, can yield invalid rationales [145]. In the context of sequential data, naive masking strategies and poorly chosen baselines may introduce artifacts that further undermine faithfulness [145].

However, even in cases where concerns about reduced accuracy with interpretable models can be addressed, intrinsically interpretable models remain infrequent in application. Recent empirical evidence on XAI in healthcare indicates that most researchers still use post-hoc rather than ante-hoc explanations [20, 64]. For example, a review by Aziz et al. [20] of 68 studies on explainable clinical decision support systems reports that nearly all systems rely on black box models with primarily model-agnostic post-hoc explanation techniques such as SHAP and LIME (see Section 4.1.2), while intrinsically interpretable models are rare.

According to Rudin [128], constructing intrinsically interpretable models often requires substantial resources, including intensive computation and deep domain expertise. In addition, commercial incentives to protect proprietary black-box predictors further reinforce the preference for post-hoc explanatory methods in both research and practice.

Interpretability vs. Accuracy One prominent dimension of the ante-hoc versus post-hoc debate concerns whether there is a necessary tradeoff between interpretability and accuracy. Some maintain that this tradeoff is unavoidable [62, 66, 67]. Others argue that for structured data, interpretable models can match the performance of more complex alternatives, and that transparency may actually support accuracy by facilitating targeted model refinement [129].

Interpretability vs. Completeness A related tension lies between interpretability and completeness. Oversimplified explanations can be misleading or nonsensical, such as saliency maps that fail to clarify how localized evidence was operationalized, whereas explanations that pursue full faithfulness and detail may overwhelm users, causing information overload [119, 128].

A pragmatic approach suggests providing explanations at different levels of detail, allowing users to adjust the depth of information to their needs and, where there is uncertainty, favoring more detailed over abbreviated descriptions [57]. In this case, explanation approaches should be assessed across the full range of this tradeoff [57]. Domain-specific strategies can also enhance specificity without inducing overload, such as supplementing prototype-based recognition with concise, quantitative textual descriptors of salient attributes [108].

4.3.3 Failure Modes for Explanations

Bove et al. [28] provide a systematic analysis of failure modes in XAI systems, distinguishing between system-specific and user-specific issues, which informs the following overview.

System-Specific Failures. System-specific failures most frequently arise in post-hoc explanation methods and often take the form of **misleading** or **contradictory** explanations. When a ML model produces inaccurate predictions, the accompanying explanations can inadvertently encourage unwarranted trust in incorrect outputs. Additionally, some explanation techniques may lack faithfulness and fail to accurately reflect the model's internal reasoning, resulting in explanations that do not correspond to the true decision-making process.

Contradictions may be introduced through conflicting elements within a single explanation, instability upon repeated evaluation of identical inputs, or disagreements between different explanation methods applied to the same decision. Such inconsistencies can confuse users, erode trust, and create uncertainty about which explanation should be regarded as reliable. In particular, visual attributions often disagree between methods and primarily indicate where a model focused attention without clarifying how evidence was used, which limits their value as standalone explanations in safety-critical domains [129].

User-Specific Failures. User-specific failures are equally consequential and affect both post-hoc and ante-hoc approaches. When explanations do not align with users' expectations or informational needs, they may lead to confusion or dissatisfaction. For example, if a user expects a straightforward feature importance explanation but instead receives a complex counterfactual account, a **mismatch** between the explanation and the user's requirements can arise.

Explanations that conflict with users' domain knowledge or intuitions can undermine trust, even when the model's predictions are accurate, resulting in **counterintuitive explanations**.

Cognitive biases also substantially affect how users interpret explanations. Non-expert users in particular may overemphasize certain elements due to individual biases or insufficient domain knowledge, which can lead to either overtrust or undertrust in the model's outputs and ultimately result in **biased inferences**.

4.3.4 Summary

Together, these findings highlight the complex challenges involved in designing effective XAI systems. Successful XAI design depends on user-centered, purpose-driven explanations that promote understanding, convey uncertainty, and foster calibrated trust. Explanations should be aligned with specific roles and tasks, use appropriate domain language, and, where relevant, adopt a conversational tone. Excessive transparency that overwhelms users and oversimplifications that prioritize persuasion over faithfulness should both be avoided. Also, there is no one-size-fits-all solution. Methods must be directly linked to explicit goals such as understandability, contestability, or recourse and must address clinical requirements.

Furthermore, ongoing tradeoffs persist between interpretability, accuracy, and complete-

ness, alongside debates over ante-hoc versus post-hoc strategies and the communication of confidence and uncertainty.

Finally, design should prepare for common failure modes, such as unfaithful or contradictory explanations and mismatches with user expectations or cognitive biases, as these factors can erode trust and compromise safe, informed use.

4.4 Evaluation of XAI and CDSSs

This section provides an overview of key guidelines and established frameworks for evaluating explanations. It begins by examining the criteria that characterize a good explanation in Section 4.4.1, then outlines central guidelines for explanation evaluation in Section 4.4.2. The evaluation frameworks are introduced and discussed in Section 4.4.3, with a summary presented in Section 4.4.4.

4.4.1 What Makes a Good Explanation?

Ongoing debates over fundamental terminologies and concepts (Section 3.2) as well as the design of XAI systems (Section 4.3.2) are paralleled by persistent disagreements in the literature regarding how to define and evaluate a *good explanation* [17].

Before this can be addressed, it is important to clarify what constitutes an explanation, as there is no universal agreement on this point either [141]. The question has deep philosophical roots. As summarized by Miller [100], an explanation is fundamentally the assignment of causal responsibility, encompassing both the process of inferring causes and the act of communicating those causes to another person. This causal and communicative perspective distinguishes explanation from related concepts such as attribution or interpretability and provides a necessary foundation for evaluating the quality, utility, or appropriateness of explanations in XAI.

Although not all explanations are causal, as Miller [100] notes, with some being non-causal (for instance, explanations prompted by the question “what happened?”), these are intentionally set aside here following the same rationale, as the focus on causal explanations reflects the predominant view within the XAI community at present.

4.4.2 Central Guidelines

A recent survey of explainable CDSSs by Aziz et al. [20] indicates that there are currently no universally accepted measures or standardized criteria for evaluating the effectiveness of XAI methods. Nonetheless, there is increasing consensus within the XAI community that the ultimate measure of an explanation’s value is its effect on human understanding and decision making [28, 100]. This user-centered emphasis reflects the underlying desiderata and design principles for explanations (see Section 4.3.1), all of which highlight the importance of user needs and context-dependent effectiveness.

The XAI agenda of the Defense Advanced Research Projects Agency further advanced

this perspective by prioritizing measurements related to user satisfaction, mental models, task performance, and appropriate trust. This approach frames explanations as tools for improved human reasoning and oversight rather than as ends in themselves [62].

Two principal approaches are commonly used to evaluate explanation quality. The first focuses on assessing **objective criteria**, such as directly measuring task performance with and without explanations [63] or quantifiable properties of the explanation, such as simplicity [68], as proxies for quality. The second centers on **subjective user perception**, which evaluates how useful or satisfactory users find an explanation [68].

Subjective perceptions are typically measured with questionnaires, yet there is no widely agreed standard for which constructs should be evaluated, such as trust, confidence, plausibility, or related aspects. Despite this lack of consensus, instruments like the System Causability Scale [68] (see Section 4.4.3) have been introduced to bring more consistency to such assessments.

When researchers rely on self-reported trust as a proxy for whether users actually rely on or override an AI system, the findings can be misleading, as stated trust or distrust often diverges from real-world behavior [49]. For example, people might prefer simpler constructs, yet perform better with more complex ones [30]. This so-called perception-behavior gap weakens the validity of trust as a measure of appropriate use [49]. In addition, cognitive biases can lead users to place too much confidence in AI systems without sufficient scrutiny (automation bias), or conversely to reject AI advice they would accept from a human (algorithmic aversion) [79]. Such biases can distort perception and decision making and accordingly must be considered when assessing explanation effectiveness.

These issues underscore the importance of combining subjective user surveys with objective measurements when evaluating the effectiveness of explanations [24, 49].

To achieve this, three common levels of evaluation are recognized: functionally-grounded, human-grounded, and application-grounded evaluation, each of which is introduced in the following section (see Section 4.4.3).

4.4.3 Evaluation Frameworks

Type of Evaluation Approaches. A widely adopted framework distinguishes among three complementary approaches: functionally-grounded, human-grounded, and application-grounded evaluation, each balancing feasibility and ecological validity [36].

Functionally-grounded evaluation relies on computational metrics as proxies for explanation quality without the involvement of human participants. Researchers define quantitative criteria such as fidelity, which measures how closely a simplified explanation model replicates the original model's behavior in a local context, sparsity, which refers to the use of a minimal subset of features to support interpretability, and consistency, which captures the stability of explanations across similar cases. This approach ensures efficiency and objectivity, enabling rapid comparison of techniques through simulation. However,

high performance on such proxy metrics does not always correspond to genuine utility or comprehensibility for end users. Functionally-grounded evaluation is most appropriate during early developmental phases or for benchmarking, but cannot substitute for human-centered assessment when the aim is to establish practical value.

In contrast, **human-grounded evaluation** engages human participants in simplified, controlled tasks that are designed to abstract from the full complexity of real-world deployment. Rather than replicating actual operational environments, these experiments isolate specific aspects of explanation quality. For example, non-expert participants may be presented with AI decisions, with or without accompanying explanations, and asked to answer questions that probe their understanding, trust, or ability to anticipate the model's behavior. Such studies, typically conducted in laboratory settings or online platforms, are less costly and time-consuming than application-grounded evaluations, but retain the important benefit of involving human judgment. Commonly used metrics include user satisfaction, perceived trustworthiness, and the ability to simulate the model's logic, which together help gauge the impact of explanations.

Application-grounded evaluation offers the most direct and contextually meaningful assessment by involving domain experts who perform authentic tasks in realistic settings. For instance, clinicians interacting with an explainable AI-based decision support tool within simulated or actual clinical workflows provide a basis for assessing explanation effectiveness using tangible outcomes such as diagnostic accuracy, error detection rates, decision speed, or user confidence. Although this approach delivers the most relevant insights into real-world utility, it is also the most resource-intensive, demanding considerable time, effort, and access to domain specialists. Application-grounded evaluations are therefore principally used in high-stakes domains, such as healthcare or finance, where demonstrating real practical benefits is essential.

Comparison of Evaluation Approaches. Because functionally-grounded evaluation depends exclusively on proxy metrics, it can misrepresent the actual usefulness for humans. High fidelity or sparsity does not guarantee that users will understand or benefit from an explanation [30]. For this reason, human-centered evaluation is widely recognized as indispensable. User studies in this context are regarded as the gold standard for assessing the quality and utility of explanations provided by AI systems [63]. However, these studies also face challenges. Many suffer from methodological limitations, such as suboptimal experimental design, which may undermine the reliability of their findings [126]. The frequent use of lay participants, rather than actual end users, raises concerns about the generalizability of results to the intended domains [13]. Furthermore, Buçinca et al. [30] observed that many user studies rely on proxy tasks or subjective measures of trust and preference, rather than measuring the joint performance of humans and AI. They conclude that only application-grounded evaluations involving professionals performing realistic tasks provide reliable insights. Importantly, task performance should be measured using domain-

specific, real-world metrics rather than only standardized assessments, as overreliance on the latter can lead to misleading conclusions [13].

System Causability Scale (SCS). Modeled after the well-established System Usability Scale (SUS) [68], the System Causability Scale (SCS) was developed by Holzinger et al. [68] as a concise Likert-based questionnaire [83] to evaluate how effectively an explanation or explainable system achieves its intended objectives. The SCS provides a systematic assessment of key qualities such as clarity, adaptability, usefulness, relevance, efficiency, and consistency, with particular emphasis on causal insights. It results in a standardized score that reflects the extent to which an explanation enhances user understanding and supports effective decision making.

4.4.4 Summary

Effective evaluation of XAI emphasizes human outcomes by prioritizing user understanding, task performance, and appropriately calibrated trust. It integrates objective measures with user perceptions and takes into account cognitive biases and the perception-behavior gap. Three complementary approaches, functionally-grounded, human-grounded, and application-grounded, balance feasibility and ecological validity. Among these, application-grounded studies yield the most relevant insights for real-world implementation in high stakes domains. To enhance the consistency of subjective assessments, standardized instruments such as the System Causability Scale enable comparison of clarity, relevance, and usefulness across different explanation methods.

4.5 Explainability in Regulatory Standards for Clinical Decision Support Systems

This section reviews regulatory frameworks that define requirements for AI-based clinical decision support systems (CDSSs) in healthcare with a particular focus on requirements for explainability. Section 4.5.1 considers the technology-neutral regulations from the United States, established by the Food and Drug Administration (FDA), and from the European Union, under the Medical Device Regulation (MDR). Sections 4.5.2 and 4.5.3 then describe how these are complemented by EU-specific measures, in particular addressing Article 22 of the General Data Protection Regulation (GDPR) and the AI Act respectively. These are followed by AI-specific recommendations from the FDA (Section 4.5.4) and the HTI-1 Rule issued by the United States Department of Health and Human Services (HHS; Section 4.5.5), both of which impose diverse expectations and requirements on AI systems in clinical practice. Finally, Section 4.5.6 provides a summary and conclusion.

4.5.1 United States FDA and EU MDR: Technology-Neutral with AI-Relevant Additions

Both the United States Food and Drug Administration (FDA) and the European Union Medical Device Regulation (MDR) govern AI-enabled tools within their established medical device frameworks rather than as a distinct AI category.

Under the United States FDA, AI-based healthcare software is typically regulated as device software functions, most often as Software as a Medical Device (SaMD), and follows the same risk-based pathways as traditional software [149]. The additional AI- and ML-specific elements appear mainly at the policy and guidance level. These include the Predetermined Change Control Plan (PCCP) for ML-enabled devices, which aims to define acceptable post-market model updates in advance [21], and the Clinical Decision Support (CDS) guidance, which clarifies when software is considered a non-device CDS [149].

In the European Union, AI-based healthcare software is treated as Medical Device Software (MDSW) within the technology-neutral legal structure of the MDR. Rule 11 in Annex VIII results in many diagnostic and therapeutic decision support applications being assigned to higher risk classes [1, 149], while Annex I sets broad requirements for software lifecycle, risk management, verification and validation, and cybersecurity [1].

Elements most pertinent to the explainability of AI-based healthcare software and in particular CDSSs are instead addressed in other regulations and guidelines, which are discussed in the following sections.

4.5.2 European Union General Data Protection Regulation: Right to Explanation (Article 22)

The General Data Protection Regulation (GDPR), implemented in 2018, serves as the foundation of the European Union's data protection framework [2]. Article 22, titled *Automated individual decision-making, including profiling*, often referred to as the *Right to Explanation*, prohibits solely automated decisions that have significant effects on individuals, except in narrowly defined situations with explicit safeguards such as human intervention and additional protections. This provision is frequently understood as a substantial endorsement of explainable XAI within the GDPR context [45, 47, 71, 142].

Nonetheless, the practical feasibility of this requirement with respect to XAI has been questioned, as for example by Hamon et al. [66], a group of researchers associated with the European Commission. They argue that, for many state-of-the-art ML systems, particularly those based on deep learning, delivering human-legible and causally grounded explanations for specific outcomes remains technically challenging. Consequently, relying exclusively on explanations may be insufficient to satisfy GDPR safeguards in certain high-risk settings. Instead, they advocate for a multifaceted approach, integrating the most feasible approaches to explainability with Data Protection Impact Assessments (DPIAs), algorithmic audits, and justification tests aligned with GDPR Article 5 principles of fairness,

lawfulness, accuracy, purpose limitation, data minimization, integrity and confidentiality, and accountability. Where explanations alone are inadequate, controllers should strengthen governance, monitoring, bias assessment, and human-in-the-loop procedures to achieve trustworthiness by design, striking a balance among performance, robustness, safety, fairness, and transparency. In this view, the authors recast the *right to explanation* as one element within a broader framework of accountability, rather than as a stand-alone technical obligation. In a subsequent publication, the authors clarify that Article 22 and related provisions on automated decision-making do not explicitly mandate the application of any specific interpretability technique for explaining decisions produced by automated systems [114].

4.5.3 European Union AI Act: High-Risk Systems and Explainable Clinical Decision Support Systems

The AI Act is the European Union's first comprehensive legislative framework for AI governance and establishes the primary legal structure for AI-based systems throughout the European Union [3]. Passed in 2024, it is set to come fully into effect in 2026.

Under the AI Act, clinical decision support systems, which are classified as medical devices requiring third-party conformity assessment, are designated as high-risk. Any AI component within such a system is similarly considered high-risk and must comply with a set of requirements that include transparency, human oversight, and traceability.

The AI Act obliges high-risk clinical AI systems to provide clinicians with clear instructions for interpreting and overseeing system outputs, to document both the intended purpose and limitations, and to maintain comprehensive logs that support traceability. Continuous risk management, sound data governance, and appropriate standards of accuracy and security are required, but the extent of explanation and human oversight may be adapted according to the specific risk and clinical context. In practice, compliant systems must supply accessible interpretation guidance, features supporting effective oversight, and auditable documentation as part of a coherent record.

Thus, the AI Act establishes functional transparency and human oversight for deployers such as clinicians and providers as mandatory obligations, treating explainability as a proportional means to ensure safe and responsible use rather than imposing a universal requirement to disclose internal model details. It does not prescribe specific XAI techniques or demand complete model transparency, nor does it prohibit the use of black box models. The focus remains on appropriateness [114].

4.5.4 FDA Guidelines: Recommendations for Meaningful, Audience-Appropriate Transparency in FDA-Regulated ML-Enabled Software

In 2024, the FDA, in collaboration with Health Canada and the UK Medicines and Healthcare Products Regulatory Agency (MHRA), issued nonbinding guiding principles

on transparency for Machine Learning-Enabled Medical Devices (MLMDs) [42]. These principles build on earlier guidance from the Good Machine Learning Practice (GMLP) in 2021 [41], spanning the entire product lifecycle. Although these documents do not establish enforceable requirements, they articulate broad expectations for meaningful, audience-appropriate transparency in FDA-regulated ML-enabled software. They establish frameworks for labeling, usability, and lifecycle controls. Importantly, the guidance does not prescribe or require specific explainability techniques. Instead, it presents explainability as a facet of transparency, outlining the characteristics of effective, fit-for-purpose disclosure and leaving space for further standards development, for example through the International Medical Device Regulators Forum (IMDRF).

The documents from the FDA, Health Canada, and MHRA clearly differentiate between recommendations and binding requirements. They emphasize principles and good practices that promote user-facing transparency and ongoing monitoring throughout the lifecycle of MLMD, rather than mandating the disclosure of model internals. In this context, explainability is considered a subset of transparency, involving clarification of the reasoning or basis for an output when such explanations are feasible and understandable to users. The central aim is to provide essential and clear information that enables intended users to interpret, trust, and act responsibly on outputs, as highlighted in GMLP Principle 9, with explicit attention to the combined performance of human-AI teams (Principle 7).

For practical implementation, a transparent MLMD should clearly state its intended purpose, target user populations, operational context, expected inputs and outputs, and anticipated impact on clinical decisions. Developers are encouraged to disclose relevant information, including clinical performance metrics, limitations, uncertainty, and confidence intervals, as well as benefits, risks, recognized biases or failure modes, gaps in available data, concise accounts of development, intended clinical validation, and lifecycle management and monitoring plans. Crucially, this information should be presented in a human-centered manner, at appropriate touchpoints throughout the product lifecycle, whether through the user interface, labeling, or training materials. It should also address points in workflows where XAI is particularly relevant, such as at key triggers or high-risk steps.

The principles further encourage developers to use datasets that represent diverse patient populations, conduct rigorous and clinically relevant testing and validation, maintain strong post-deployment monitoring, and integrate quality and security engineering throughout development and operational phases.

In summary, the transparency and GMLP principles from the FDA establish deployer- and user-centered transparency as the recommended baseline for MLMDs. They call for accessible information regarding purpose, limitations, performance, and lifecycle management, support for safe human oversight and decision-making, and assurance that models are auditable and subject to continuous monitoring. These guiding documents do not require specific XAI methods or comprehensive model disclosure. Legally binding obligations for any particular product are determined by the device's FDA regulatory pathway, including

requirements related to labeling, validation evidence, quality management, and applicable standards. The guiding principles thus offer a framework for effective transparency and sound ML practices in the current regulatory landscape.

4.5.5 HTI-1 Rule from the United States Department of Health and Human Services: Predictive Decision Support Interventions

In 2024, the Health and Human Services (HHS) introduced the *Health Data, Technology, and Interoperability: Certification Program Updates, Algorithm Transparency, and Information Sharing Rule* (HTI-1 Rule), establishing federal requirements for Predictive Decision Support Interventions (DSIs) [5]. Predictive DSIs are defined as "[...] technology that supports decision-making based on algorithms or models that derive relationships from training data and then produces an output that results in prediction, classification, recommendation, evaluation, or analysis." [5]. This definition places Predictive DSIs within the broader category of AI-based CDSSs.

HTI-1 is an Office of the National Coordinator for Health Information Technology (ONC) final rule, updating the ONC Health IT Certification Program. It sets requirements for developers of certified health IT, that is, those creating or supplying Electronic Health Record (EHR) solutions, with respect to their certified Health IT Modules. The rule does not confer approval on individual clinical tools. Instead, when an EHR vendor provides a predictive DSI as part of a certified module, HTI-1 requires that transparency and risk-management capabilities be built into that module. Provider organizations, such as clinics, retain the authority to determine which tools to deploy, and any relevant FDA obligations continue to apply. The practical implications of EHR vendor solutions will be examined in greater detail in Section 5.3.

The rule requires that certified health IT systems supplying Predictive DSIs ensure deployer-facing transparency by granting certain users access to *source attributes*, defined as "categories of technical performance and quality information" [5]. These attributes encompass essential factors such as purpose, development, limitations, population relevance, fairness, validation status, performance, and maintenance. Certified modules must enable users not only to view but also to manage and update this information in plain language. This framework ensures that key facts concerning the quality and risk associated with Predictive DSIs are documented, reviewable, and auditable, thereby supporting informed use and effective oversight.

Explainability is not mandated as a specific algorithmic method. Rather, HTI-1 incorporates the principle of *intelligibility* within the required Intervention Risk Management (IRM) process. For each Predictive DSI supplied, developers are required to assess risks and potential adverse impacts across the following eight characteristics [5]: validity, reliability, robustness, fairness, intelligibility, safety, security, and privacy. Developers must implement mitigation practices and maintain governance policies addressing data

acquisition, management, and use.

In summary, HTI-1 establishes mandatory functional transparency and oversight guidance through source attributes and IRM, treats explainability as a proportional obligation to provide intelligibility for the intended audience, and emphasizes disclosure of purpose, limitations, performance, fairness, and maintenance over exposure of model internals or prescription of particular XAI methods. Related considerations, including bias, oversight, safety, security, and privacy, are addressed through the requirements for source attributes and the IRM framework, both of which must be published and kept up to date.

4.5.6 Summary and Discussion

The regulatory landscape for explainable clinical decision support systems distinguishes clearly between mandatory requirements and voluntary recommendations. Binding obligations arise primarily from the European Union's AI Act for high-risk clinical AI systems and the HTI-1 Rule for certified health IT modules, both of which require functional transparency, mechanisms for human oversight, and auditable documentation. By contrast, FDA guidance through the GMLP principles and international standards such as those from the IMDRF offer recommendations regarding transparency and explainability but do not prescribe specific XAI techniques.

Regulatory frameworks consistently prioritize intelligibility and traceability for deployers over disclosure of model internals. The emphasis is on enabling clinicians to comprehend the system's purpose, limitations, and appropriate contexts of use, rather than on specific technical approaches to explainability. This approach seeks to balance the need for transparency with the operational realities of advanced AI-based systems, supporting accountability through documentation, monitoring, and human oversight rather than mandating particular forms of explainability.

5 Practical Landscape of Explainable Clinical Decision Support

This section reviews the practical landscape surrounding XAI and CDSSs. It first addresses contradictions and recent trends, focusing on user studies on explainable CDSSs, as presented in Section 5.1. Section 5.2 discusses commonly cited obstacles to the adoption of AI-enabled CDSSs. Finally, Section 5.3, assesses the impact and significance of XAI in clinical practice.

5.1 Findings from Practice: Contradictions and Emerging Trends in Explainable Clinical Decision Support

This section synthesizes insights from observations of AI-based systems in practice, with a primary focus on user studies involving explainable CDSSs. The discussion is organized through two complementary perspectives. Section 5.1.1 examines problematic, unexpected,

and contradictory findings that challenge prevailing assumptions regarding explanations and trust. Section 5.1.2 then draws out common patterns that signal trends in confidence communication and interaction design.

5.1.1 Problematic, Unexpected and Contradictory Findings

Advice Effects and the Perception-Behavior Gap. In a large, application-grounded study by Gaube et al. [49] on physician diagnoses based on chest radiographs, participants often rated advice from AI sources as lower quality, a phenomenon known as algorithmic aversion, yet still relied on it as much as human-labeled advice. This reveals a perception-behavior gap between clinicians' self-reported trust and their actual choices at decision time. This gap is problematic because excessive trust in incorrect advice undermines accuracy, while insufficient trust in correct advice can have similar negative effects [127]. In contrast, in a follow-up study using the same material, Gaube et al. [50] were surprised, reporting that ratings of advice quality showed only minimal differences between AI and human sources, indicating an absence of clear algorithmic aversion or appreciation.

Taken together, these findings confirm that the perception-behavior gap, where clinicians' stated trust in advice diverges from their actual reliance, is a recognized and persistent issue (see Section 4.4.2). However, it remains unclear under which specific conditions this gap emerges or can be mitigated. This uncertainty highlights the importance of evaluation frameworks and deployment strategies that are not solely based on self-reported measures but include objective measures like collaboration performance or accept and override rates (see Section 4.4.2).

Explanation Effect on Trust. Empirical evidence on whether explanations increase trust remains mixed. In intensive care settings, interactive transfactual explanations (see Section 4.1.4) increased clinicians' trust and substantially aligned clinicians' mental models of feature relevance with model behavior [71]. By contrast, broader syntheses report studies showing trust increases, null effects, and findings where explanations either enhance or diminish trust, depending on user and context [127].

A study on vignette-based consent by Baron et al. [24] indicates that any explanation type, causal, counterfactual, or story-based, tends to outperform none in terms of subjective measures like helpfulness or trust, but the differences among explanation formats are small. Clinical stakes, rather than explanation format, more strongly influenced outcomes and attitudes, with greater caution exercised under high-stakes conditions regardless of whether advice was attributed to an AI system or a physician.

In sum, it remains unclear what consistently drives trust in clinical AI explanations: sometimes explanation type and interactivity have noticeable effects, but at other times context, such as clinical stakes or user expectations, matters far more, and explanation format makes little difference. These findings highlight persistent gaps in understanding when and how explanations meaningfully affect trust.

Contradictory Findings Beyond Clinical Settings. Outside clinical domains, user studies reveal additional contradictions that challenge assumptions regarding the effectiveness of explanations.

Bansal et al. [23] reports findings from prior studies that gains in human-AI team performance emerge primarily when the AI system alone outperforms both the best-performing individual and the best-performing team, implying explanations can at times mislead rather than assist users.

Similarly, a user study by Poursabzi-Sangdeh et al. [119] finds that providing explanations can inadvertently decrease users' ability to detect prediction errors, attributed to information overload, while, in contrast, a user study by Kulesza et al. [78] found that detailed explanations were generally preferred and more helpful, whereas sparse or overly simple justifications were met with dissatisfaction.

Moreover, Schmidt and Biessmann [131] found that transparency accompanying uncertain or incorrect predictions can introduce algorithmic bias, especially among risk-averse individuals, as explanations may unwittingly reinforce biased inferences or mislead users, rather than improving understanding of model limitations.

Taken together, these findings underscore the complexity of designing effective XAI systems and the necessity of carefully calibrating the granularity of explanations. They further illustrate that the interpretability-completeness tradeoff, although well established (see Section 4.3.2), manifests differently depending on practical context, making it essential to adjust and assess explanation detail based on situational needs. This underscores the necessity of thoughtful user interaction design and sustained attention to potential biases that can emerge during AI-assisted decision making.

5.1.2 Common Findings and Trends

Confidence Drives Acceptance while Transparency Modulates Borderline Cases. Patterns of acceptance and override closely track calibrated confidence. In a study by Yu et al. [159], recommendations with very high confidence (90–99%) were rarely overridden (approximately 1.7%), whereas those with low confidence (70–79%) were almost always overridden (approximately 99.3%). Transparency primarily influenced decisions in mid-confidence ranges, and the combination of high explainability with high confidence yielded the lowest override rates [159]. Displaying high confidence was found to induce over-reliance, while low confidence generally reduced trust and prolonged decision making [159]. In general, clinicians prefer advice that communicates uncertainty [49]

These findings highlight a strategic opportunity for XAI design. Systems that clearly communicate both confidence and limitations can help mitigate over-reliance, particularly when clinicians feel uncertain [49]. Rather than simply presenting confidence scores, systems can use moments of physician uncertainty as natural intervention points, offering calibrated explanations that support more informed decisions about when to trust or

override recommendations.

The broader operational insight is that calibrated confidence serves as the primary determinant of reliance, with explanations acting as tie-breakers in ambiguous situations, and strategic confidence communication functioning as a safeguard against over-reliance.

On-demand advice, cognitive forcing, and progressive disclosure. Providing advice and explanations on demand, rather than by default, can reduce automation bias by limiting anchoring on incorrect suggestions [49]. Interface-level cognitive forcing functions that prompt analytical checking can decrease over-reliance more effectively than explanations alone, though they may introduce trade-offs in subjective user preference [31]. Progressive disclosure and role-specific content help to avoid information overload and increase the likelihood that explanations are accessed when most relevant [12, 127]. Conversely, lengthy or redundant explanations are frequently skipped and raise cognitive load, which diminishes use in practice [127].

These observations emphasize that interaction design itself, not just the content of explanations, functions as a key safety mechanism in human-AI teaming.

5.2 Adoption Barriers of AI-Enabeld CDSSs

Although healthcare is generally seeing increasing approval rates of AI-based systems, as will be discussed in Section 5.3.1, the integration of AI-enabled CDSSs into clinical practice is impeded by several well-documented barriers. This section presents a structured summary of frequently expressed challenges using a four-level taxonomy: user level (Section 5.2.1), data level (Section 5.2.2), research and development level (Section 5.2.3), and operational level (Section 5.2.4). At each level, challenges are examined through four primary impact categories: trust and adoption, which concerns user acceptance and confidence, technical performance, which addresses system capabilities and data quality, workflow integration, which focuses on incorporation into clinical routines, and sustainability, which considers long-term viability and safe ongoing use. Section 5.2.5 discusses the cross-cutting nature of the impact categories and Table 1 provides an overview of these implementation challenges by level and impact category.

5.2.1 User Level

At the user level, obstacles arise across several aspects of implementation.

Trust & Adoption. Clinicians frequently express skepticism toward these systems, stemming from challenges in understanding how decisions are produced, which raises concerns about the accuracy and reliability of AI-driven recommendations [69, 138]. The prevalence of excessive or irrelevant alerts contributes to alert fatigue, prompting clinicians to disregard or mistrust notifications [51, 143]. High override rates further erode clinician confidence in AI systems and impede adoption, as frequent interventions increase workload,

Table 1: **Implementation challenges for CDSSs**, organized by level and impact category. Sections 5.2.1 through 5.2.4 provide detailed explanations for each level. Color coding is used to facilitate navigation of the impact categories across levels.

Level	Key Challenges (by Impact Category)	
User	Trust & Adoption:	Clinician skepticism, alert fatigue, override rates, job displacement concerns
	Technical Performance:	Outcome limitations, perceived utility requirements, technical proficiency barriers
	Workflow Integration:	Workflow disruption, poor UI/UX, learning curve burden
	Sustainability:	Over-reliance risk, skill degradation
Data	Trust & Adoption:	Healthcare data inequity concerns, biased outcomes
	Technical Performance:	Data quality issues, external data dependencies, data access barriers
	Workflow Integration:	Inadequate system design, standardization gaps
R&D	Technical Performance:	Limited public datasets, XAI challenges, interdisciplinary co-design
Operational	Technical Performance:	Interoperability challenges, heterogeneous data sources
	Workflow Integration:	EHR coding inconsistencies, FHIR complexity, legacy system challenges
	Sustainability:	System maintenance, regulatory compliance, financial constraints

contribute to cognitive fatigue, and result in inconsistent clinical decisions [159]. In addition, apprehension over potential job displacement by AI leads to resistance among clinicians [113, 131].

Technical Performance. CDSSs often do not demonstrate consistent improvements in clinical outcomes [49]. The willingness of clinicians to use prediction tools depends on perceived utility and technological familiarity [55]. Effectiveness is affected by users' technical proficiency, with overly complex systems or those that deviate from established workflows presenting barriers to engagement [143].

Workflow Integration. Inadequate integration with existing workflows can disrupt clinical routines, increase cognitive burden, and limit the time available for patient care [143, 153]. The absence of intuitive, user-centered interfaces often impedes adoption [69]. The perceived responsibility and initial learning curve associated with adopting new AI tools may also discourage clinicians from embracing these technologies [131].

Sustainability. Excessive reliance on CDSSs can produce a carryover effect, where users' independent clinical skills may decline over time [143].

5.2.2 Data Level

The data level includes challenges related to data quality, availability, and equity that directly influence system performance.

Trust & Adoption. There is mounting concern that AI-driven systems may amplify

existing inequities in healthcare data, which can produce biased outcomes [110, 131].

Technical Performance. The performance of CDSSs is highly dependent on the quality and availability of data, which vary widely across healthcare settings [143, 153]. Reliance on external and dynamic data sources can create operational challenges, such as outdated medication lists, that degrade decision support [143]. Many healthcare organizations remain reluctant to share sensitive patient data due to legal or ethical concerns, creating barriers to data integration and limiting the effectiveness of systems that require comprehensive patient information [11, 143].

Workflow Integration. Inadequate system design can lead users to enter generic or incorrect data, compromising decision support quality and emphasizing the need for standardized information systems [143].

5.2.3 Research and Development Level

At the research and development level, substantial barriers impede the advancement and validation of AI-enabeld CDSSs.

Technical Performance. Limited access to public datasets remains a significant barrier. As noted by Aziz et al. [20], more than 55% of tabular datasets used in explainable CDSSs studies in healthcare are private, which limits the potential for replication and comparative analysis. Further XAI challenges include the scarcity of practical tools, the multidimensional nature of explainability, and the pressing need for rigorous evaluation, including both automated and user-centered approaches [12]. Frequently, studies rely on only one XAI method, which can constrain the scope of interpretability assessment [20]. The field is also characterized by ongoing discussions around design tradeoffs (see Section 4.3.2), established failure modes of explanations (Section 4.3.3), and the complex task of evaluating XAI systems (Section 4.4.1). Evidence from practice remains contradictory and fragmented, without systematic synthesis into actionable guidance for different contexts (Section 5.1). Also, explanation is inherently a social and user-dependent process so that effective explainability requires co-design with clinicians, patients, and experts in the social and behavioural sciences, philosophy, psychology, and cognitive science, to ensure explanation practices are aligned with users' expectations and needs [17, 100].

5.2.4 Operational Level

Operational-level challenges further complicate the deployment and ongoing viability of CDSSs.

Technical Performance. Persistent issues of interoperability and data transportability confront CDSSs, which must manage heterogeneous clinical data sources [143].

Workflow Integration. The adoption of the HL7 FHIR standard has advanced integration and interoperability among healthcare information systems (see Section 5.3). Nevertheless, substantial barriers remain, notably the ongoing inconsistencies in clinical concept coding

within EHR data, which hinder accurate exchange and interpretation of information [143]. Also, the implementation of FHIR can be complex, especially for legacy systems, and may require customization to suit specific organizational needs [124].

Sustainability. Maintenance of both system and content, involving technical upkeep and continual updates of knowledge bases to reflect evolving medical guidelines, is essential but is often overlooked [143]. Regulatory compliance and legal issues are major hurdles for responsible system deployment [153] and are further complicated by the evolving regulatory environment for AI (see Section 4.5). Financial constraints, including initial investments, ongoing training, and continual system updates, add further challenges to widespread adoption and the sustained effectiveness of CDSSs [143].

5.2.5 Cross-Cutting Themes.

A number of challenges extend across multiple organizational levels, emphasizing the interconnectedness of CDSSs implementation. Barriers to trust and adoption are evident at both the user level, with clinician skepticism and alert fatigue, and the data level, with concerns about equity. This suggests that transparency and fairness are essential across the entire system. Technical performance issues appear at every level, encompassing user proficiency, data quality, research dataset limitations, and operational interoperability. Workflow integration challenges are similarly pervasive, affecting user experience, data standardization, and compliance with operational standards. Concerns about sustainability arise at the user and operational levels and include the risks of individual skill deterioration and organizational burdens related to finance and maintenance. These cross-cutting themes highlight the need for holistic strategies that address implementation barriers across all dimensions of the organization.

5.3 Explainable CDSSs Systems in Practice

This section discusses the current landscape of AI-enabled medical devices in clinical practice, emphasizing systems cleared by the FDA for use in the United States. Section 5.3.1 first reviews trends among FDA-cleared AI-enabled medical devices and the technological foundations supporting their integration into clinical workflows. The subsequent analysis in Section 5.3.2 focuses on the explainability features of these systems, with particular attention to the radiology domain, which accounts for most deployments in practice.

5.3.1 AI-enabled Systems in Practice

The United States Food and Drug Administration (FDA) maintains a comprehensive registry of AI-enabled medical devices approved for clinical use [148]. Key findings from this database are summarized in Figure 2.

Figure 2A displays a substantial acceleration in the number of FDA-cleared AI-enabled medical devices beginning in 2016.

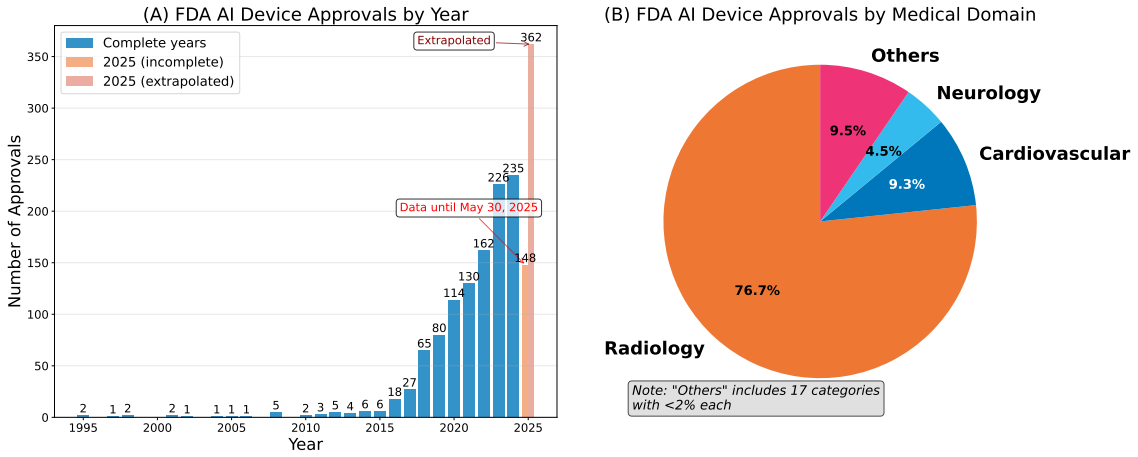


Figure 2: **Insights from the list of FDA-cleared AI-enabled medical devices** [148]**.** Panel A shows the number of devices by year, illustrating a marked increase since 2016. The 2025 bar distinguishes both approvals to date (incomplete year) and an extrapolated projection that scales the observed approval rate up to the full year based on the number of days elapsed so far, assuming a constant approval rate. Panel B shows the distribution of all devices among clinical domains. Medical device panels with less than 2% share are grouped as "Others". Radiology dominates with over three quarter of all devices.

While this list encompasses more than just CDSSs, a likely driver of the observed growth is the adoption of the Substitutable Medical Applications and Reusable Technologies on Fast Healthcare Interoperability Resources standard, commonly referred to as SMART on FHIR. Introduced by Mandel et al. [92] in 2014, SMART on FHIR is now widely used to standardize and secure the integration of third-party applications with Electronic Health Records (EHRs). In this context, FHIR refers to HL7 FHIR, a web-based data exchange standard developed by Health Level 7 (HL7). HL7 itself is a long-established framework for health information exchange that defines formats and protocols to enable the secure sharing of medical and administrative data across diverse healthcare systems. HL7 FHIR facilitates flexible, modular, and interoperable sharing of clinical information using RESTful APIs, supporting the integration of contemporary digital tools with both modern and legacy healthcare technologies [111].

SMART on FHIR enables secure apps to interact with EHR data for both clinicians and patients by extending the FHIR standard with robust authentication protocols, such as OAuth 2.0 and OpenID Connect [111].

The principal advantage offered by SMART on FHIR is enhanced interoperability. The framework allows a wide range of digital tools to access and update patient records via uniform APIs within clinical workflows, promoting electronic data exchange and reducing data silos and manual effort [124]. This supports real-time clinical decision support, analytics, and patient engagement. By encouraging open and modular health IT ecosystems, organizations can adapt solutions to specific needs without large-scale EHR replacements [92].

Leading EHR vendors that hold Office of the National Coordinator for Health Informa-

tion Technology (ONC) certification(see Secton 4.5.5), such as Epic and Cerner, together representing more than half of the United States market share [25], now provide strong support for third-party application integration through the SMART on FHIR standard [51]. As of October 2025, Epic's App Gallery features 54 CDSSs in addition to many other applications, a significant proportion of which integrate AI functionality [4].

5.3.2 Explainable AI-Enabled Systems in Practice

As shown in Figure 2B, more than three quarters of all FDA-cleared AI-enabled medical devices are in the radiology domain. In 2023, within this field, Gaube et al. [50] reported the availability of over 190 CE-marked and nearly 100 FDA-cleared AI software solutions. More recently, McNamara et al. [96] conducted a comprehensive analysis of 104 FDA-cleared computer-aided detection (CAD) systems designed to support clinicians in identifying a range of lesions in medical images, including intracerebral hemorrhage and breast cancer, across fourteen distinct lesion types.

The study identified five principal types of CAD systems. CADt (triage) systems flag cases for clinician prioritization without highlighting specific findings. CADe (detection) systems indicate lesion locations on images. CADx (diagnosis) systems provide a numeric or categorical risk score without localizing findings. CADe/x systems combine lesion detection with diagnostic scoring. CADa systems autonomously interpret cases without clinician review. These categories differ both in their outputs, which range from simple binary triage to detailed visual markers and risk scores, and in the extent to which they support or replace clinician involvement.

All CADt devices, representing 59% of the systems reviewed, offered no explanations, which aligns with their primary function of providing only a priority flag to clinicians. Most CADe, CADx, and CADe/x systems, just under forty in total, provided only basic location information, typically as bounding boxes or segmentation masks, and did not offer more granular, pixel-level feature importance. Only eleven systems presented more advanced forms of explanation, such as semantic descriptors (five devices), for example, labeling a lesion as *round*, lesion size quantification (five devices), or the display of similar case examples (one device).

In summary, the majority of AI-enabled medical devices analyzed in the radiology domain, representing the largest deployment domain for CDSSs, offer little or no explainability. Although Section 4.5 notes the absence of explicit regulatory demands for explainability, this analysis demonstrates that, in practice, the industry has tended to deprioritize explainability, making non-explanatory outputs the norm.

6 Publications

This section presents a list of the publications produced during the doctoral studies in Section 6.1 and provides a brief introduction of each work in Sections 6.2 to 6.5,

focusing on the backgrounds, main topics, and general approaches. A detailed discussion of the results, the interrelations of the publications, and their relation to the objectives of the doctoral studies is deferred to Section 7. The full manuscripts of the included publications are provided in Appendices A, C, and D.

6.1 List of Publications

Table 2 provides a summary of the publications created throughout the doctoral studies. The doctoral candidate is the primary author of all publications included in this dissertation.

Table 2: **Publications Produced During the Doctoral Studies.** The column *Included* specifies whether the respective publication is incorporated within this dissertation.

Title	Year	Publication	Included
SPARROW: Semantically Coherent Prototypes for Image Classification	2021	British Machine Vision Conference (**BMVC**) [74]	Yes
Coherence Evaluation of Visual Concepts with Objects and Language	2022	**ICLR**2022 Workshop on the Elements of Reasoning: Objects, Structure and Causality [80]	No
Comprehensive Polysomnography (CPS) Dataset: A Resource for Sleep-Related Arousal Research	2024	**PhysioNet** [75]	Yes
ALPEC: A Comprehensive Evaluation Framework and Dataset for Machine Learning-Based Arousal Detection in Clinical Practice	2025	Proceedings of the sixth Conference on Health, Inference, and Learning (**CHIL**), volume 287 of Proceedings of Machine Learning Research (**PMLR**) [77]	Yes
Assessing the Real-World Utility of Explainable AI for Arousal Diagnostics: An Application-Grounded User Study	2025	**arXiv** preprint [76]	Yes

6.2 SPARROW: Semantically Coherent Prototypes for Image Classification

This publication investigates prototype classification, a form of case-based reasoning (see Section 4.1.4), that functions as an intrinsically interpretable XAI method (see Section 4.1.2).

The work builds on the *ProtoPNet* architecture introduced by Chen et al. [32]. *ProtoPNet* combines a convolutional base network with prototype vector weights to compute similarity scores between latent space sample embeddings and prototypes, thereby providing intrinsic interpretability. Along with the approach of Li et al. [82], this architecture represents a

key advance in prototype learning within the latent spaces of deep neural networks for image classification [102].

We adopt the *ProtoPNet* framework for several reasons. Notably, Chen et al. [32] show that the architecture achieves high classification accuracy while preserving intrinsic interpretability, effectively avoiding issues surrounding faithfulness of post-hoc XAI methods (see Section 4.3.2). *ProtoPNet* uses latent representations of actual training sample image patches as prototypes, making model decisions directly traceable to the training data. This is in contrast to the approach of Li et al. [82], which relies on an autoencoder to visualize prototypes from latent representations. This distinction renders *ProtoPNet* conceptually advantageous over the latter, since it is certain that the prototypes are contained within the training data distribution.

However, our analysis of the Chen et al. [32] approach reveals persistent challenges in connecting prototypes to clear semantic meaning. Some prototypes correspond to input activations that are diffuse and lack focus on meaningful image regions, while others display overlapping or inconsistent semantics across samples. These issues diminish interpretability. To address these limitations, we developed a training procedure that incorporates additional loss terms designed to promote semantic coherence, consistency, and diversity among prototypes. Furthermore, we introduced the `SPARROW` evaluation framework to enable a functionally-grounded (see Section 4.4.3), quantitative assessment of these semantic desiderata. We evaluate our approach using the Caltech-UCSD Birds-200-2011 (CUB) dataset, a widely used benchmark for image classification that focuses on bird species [152].

6.3 Comprehensive Polysomnography (CPS) Dataset: A Resource for Sleep-Related Arousal Research

The Comprehensive Polysomnography (CPS) dataset was collected during a monocentric study conducted in collaboration with Klinikum Esslingen and NRI Medizintechnik GmbH between 2021 and 2022. The dataset encompasses 113 overnight polysomnography recordings, each featuring up to 36 raw data channels and 23 derived channels, alongside 81 distinct types of annotated events per subject. Supplementary information was gathered through diverse questionnaires and the extraction of medical diagnoses from doctor's letters of the practicing sleep medicine physician [75].

This data collection was conducted under a medical study approved by the ethics committee of the federal state, Landesärztekammer Baden-Württemberg, Germany, in October 2020, and is registered with the German Clinical Trials Register (DRKS) [156].

The medical study primarily aimed to explore the utilization of AI-based pulse wave analysis (PWA) in advancing sleep-related arousal diagnostics. Its objectives included enhancing both the quality and efficiency of diagnostic workflows performed by medical professionals through an AI-powered clinical decision support system. Furthermore, the study

investigated the potential of PWA measurements derived from the photoplethysmography modality to replace conventional electrode-based modalities such as electroencephalogram (EEG), electromyography (EMG), and electrooculography (EOG). This shift was intended to reduce technical complexity and support broader efforts aimed at enabling home-based arousal detection in ambulatory rather than clinical environments.

The Comprehensive Polysomnography (CPS) dataset has been released as a publicly available resource for sleep-related arousal research. In September 2024, it became available on the PhysioNet repository under the PhysioNet Credentialed Health Data License 1.5.0 [75] and may be accessed by credentialed users.

6.4 ALPEC: A Comprehensive Evaluation Framework and Dataset for Machine Learning-Based Arousal Detection in Clinical Practice

Analysis of the data collected for our medical study (see Section 6.3) revealed an unexpected feature: medical scorers annotated only the onsets of arousal events rather than their full durations. This approach differs from widely used polysomnography datasets, which typically provide complete event annotations, and posed challenges for conventional training methods that rely on fully labeled event boundaries.

Consultations with sleep medicine experts confirmed that this practice is widespread in clinical settings, particularly in sleep laboratories focused on patients with Obstructive Sleep Apnea sleep disorders, where only the onset of arousals is deemed clinically relevant. Omitting event offsets simplifies and accelerates the annotation process.

Imposing stricter annotation requirements would burden sleep laboratories already constrained by scorer availability and patient demand. To accommodate this clinical reality, we propose focusing on the detection of arousal onsets rather than full event boundaries. This adjustment enables the training of arousal detection models within any sleep laboratory without the need to modify established annotation workflows. We systematically examine multiple training strategies, comparing approaches that detect arousal onsets as discrete time points, fixed windows, or constructed intervals with a baseline that utilizes full event detection including both onset and offset labels. In addition to standard window-based classification, we incorporate *DeepSleep 2.0* [40, 81], a convolutional neural network (CNN)-based continuous segmentation model that performs pointwise separation of time series into arousal and non-arousal segments, accommodating an arbitrary number of data channels. We train several models using both the CPS dataset and the 2018 PhysioNet Challenge dataset [56, 58].

Next, our literature review reveals that performance evaluation methodologies for arousal detection are highly fragmented. Protocols often align with the training scheme, with window-based models evaluated in corresponding fashion and segmentation models assessed via pointwise comparison to ground truth. We find that both evaluation strategies fall short of addressing the distinctive requirements of arousal detection for clinical practice.

Moreover, we find that the performance metrics used in most existing studies fail to capture practical considerations associated with utilizing an AI-based clinical decision support systems for arousal detection in clinical practice. We tackle both these issues by introducing the ALPEC framework.

Finally, we find a lack of multimodal data usage in other works on arousal detection. Since the CPS dataset provides a rich set of modalities (see Section 6.3), we investigate the impact of using various modalities for arousal detection.

6.5 Assessing the Real-World Utility of Explainable AI for Arousal Diagnostics: An Application-Grounded User Study

The principal contribution of this dissertation is a rigorously designed application-grounded (see Section 4.4.3) user study that evaluates the real-world clinical utility of an AI-powered cdss (CDSS) for arousal detection. This study utilizes the CPS dataset (see Section 6.3) and features the highest-performing model identified in our previous work, optimized for clinical utility through the ALPEC framework (see Section 6.4). Eight professional sleep scorers, representing a range of companies and sleep laboratories, participated in a user study that examined two central aspects of the human-AI collaboration: the transparency of AI assistance (comparing black-box and transparent support) and the timing of that assistance (support provided from the outset versus targeted post-hoc quality control). This factorial design enabled a systematic assessment of the effects of explanation and workflow timing on decision quality, while benchmarking human-AI collaboration against independent human scoring is also performed.

The web-based CDSS (more specifically a diagnostic decision support system; see Section 3.3.1), developed with guidance from an expert trainer in arousal scoring to assure ecological validity and alignment with real-world workflows, supports both manual and AI-augmented scoring. In its transparent mode, the system presents calibrated confidence estimates along with post-hoc explanations drawn from both local and global feature attributions, utilizing DeepLIFT (see Section 4.1.2), tailored to the characteristics of time-series sleep data.

All participants received standardized training prior to the start of the study. The study was structured into several phases, allowing each participant to perform manual arousal scoring, re-assessment with AI assistance for quality control, and scoring with AI assistance from the outset. AI assistance was provided in either a black-box or transparent mode. Each participant completed all phases on the same subject data, allowing for a repeated measures analysis, within a counterbalanced design to mitigate order effects.

To provide a holistic assessment of the effectiveness of the CDSS, the study incorporates event-level metrics, evaluation based on total arousal event counts, measures of time efficiency, and user experience captured through standardized questionnaires and targeted queries. The experimental protocol employs a dual ground-truth strategy, using (i) the

labels from the CPS dataset and (ii) a consensus built from annotations of unaided human scorers participating in the study. The study employs robust statistical procedures to ensure rigorous and reliable evaluation of its findings.

7 Results and Discussion

This section summarizes the findings and contributions of the published works and discusses them in greater detail. For clarity and brevity, the publications are referenced using their abbreviated titles as follows.

- SPARROW: denotes the publication *SPARROW: Semantically Coherent Prototypes for Image Classification* [74] (see Appendix A).
- CPS: denotes the publication *Comprehensive Polysomnography (CPS) Dataset: A Resource for Sleep-Related Arousal Research* [75] (available online on PhysioNet).
- ALPEC: denotes the publication *ALPEC: A Comprehensive Evaluation Framework and Dataset for Machine Learning-Based Arousal Detection in Clinical Practice* [77] (see Appendix C).
- User Study (capitalized): denotes the publication *Assessing the Real-World Utility of Explainable AI for Arousal Diagnostics: An Application-Grounded User Study* [76] (see Appendix D).

Figure 3 provides an overview of the publications, highlighting their interconnections, commonalities, and distinctions which are presented in detail in the subsequent sections. Sections 7.1 to 7.3 follow a consistent structure. For each publication, the *overarching contributions and relevance of the findings* are discussed, summarizing the main contributions and their significance to the overall objectives of the doctoral studies and the broader fields when viewing each publication in isolation. This is supplemented by the *link to subsequent publications*, which outlines how the publication connects to and informs later works that includes more narrowly focused contributions. The discussion then turns to the User Study (Section 7.4), which serves as the central publication of this dissertation and adopts a distinct structure. Section 7.5 then considers issues of workflow design, compares avenues of decision support and autonomous AI decision making, and discusses future directions for arousal diagnostics and XAI. For an introduction to each publication, refer to Sections 6.2 to 6.5.

7.1 SPARROW: Semantically Coherent Prototypes for Image Classification

7.1.1 Overarching Contributions and Relevance of the Findings

This publication marks the initial methodological investigation in this doctoral research, focusing on prototype classification of image data. It delivers two primary contributions.

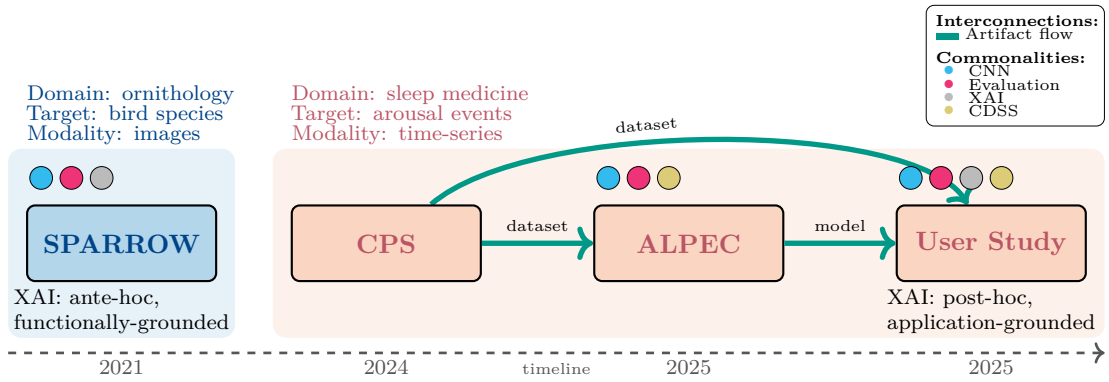


Figure 3: **Overview of the four publications**, depicting their interconnections, commonalities, and distinctions. **Interconnections** (shown by teal arrows as artifact flows) highlight that CPS provides training data for ALPEC as well as realistic data for the User Study, while ALPEC supplies an optimized model to the User Study. **Commonalities** (represented by colored circles; see legend) include a shared emphasis on convolutional neural network (CNN) usage and evaluation among SPARROW, ALPEC, and the User Study, and a focus on XAI in both SPARROW and the User Study. Additionally, ALPEC and the User Study both address aspects of clinical decision support systems (CDSSs). **Distinctions** are as follows: SPARROW targets bird species classification in images using prototype classification methods, adopts an ante-hoc explainability paradigm, and employs functionally-grounded evaluation. In contrast, CPS, ALPEC, and the User Study focus on sleep medicine, particularly on arousal event detection from time-series data. The User Study also uniquely applies post-hoc explainability together with application-grounded evaluation. For reference, the related sections are as follows: SPARROW – Overview, Results and Discussion, Manuscript; CPS – Overview, Results and Discussion; ALPEC – Overview, Results and Discussion, Manuscript; User Study – Overview, Results and Discussion, Manuscript.

First, it introduces the SPARROW framework, which enables functional-grounded evaluation of the semantic coherence of prototype classification methods. Second, it applies this framework to improve the prototype classification approach *ProtoPNet* by Chen et al. [32] (see Section 6.2).

Our framework offers a tool to be used in future research in prototype classification, and as *ProtoPNet* remains highly relevant with a multitude of recent studies utilizing or extending it, such as Pathak et al. [116], Wolf et al. [157], Zhu et al. [161], so that our methodological improvements have practical value for broader adoption. In the context of the objectives laid out for this doctoral work (see Section 2), SPARROW makes a clear contribution to the field of XAI. Specifically, it complements the subsequent studies by addressing methodological advances in an ante-hoc XAI approach for case-based reasoning applied to image data, thereby broadening the methodological scope of the doctoral research.

7.1.2 Link to Subsequent Publications

Initially, I intended to employ prototype classification for arousal detection. As a form of case-based reasoning (CBR), this explanatory approach is generally recognized as well suited to the medical domain and to human reasoning more broadly (see Section 4.1.4). *ProtoPNet*, in particular, offers a largely model-agnostic approach for deep neural networks,

promising both strong predictive performance and faithful explanations due to its intrinsic interpretability (see Section 4.3.2).

Following the initial methodological advances with SPARROW in the image domain, a transition to time series data would have required substantial adaptation. A key challenge arises from the multivariate nature of physiological signals. In contrast to image data, where feature dimensions are spatially coherent, polysomnography (PSG) datasets typically comprise numerous channels (see Section 3.4) where most represent distinct physical or physiological processes. While these challenges have now been addressed [160], suitable methods were not available at the time of this publication.

For EEG data, one possible solution involved using time-frequency transformations, allowing prototype-based methods to be applied to spectrograms that can be treated similarly to images, as demonstrated by Foroughi et al. [43]. However, this approach is unlikely to extend to most non-EEG modalities and would also present arousal scorers with unfamiliar data representations, contradicting principles of user-centered explanation design (see Section 4.3.1).

Thus, implementing *ProtoPNet* for arousal diagnostics would have necessitated a sustained focus on methodological advancement. While such a direction would have posed an interesting methodological challenge in itself, the principal aim of this doctoral research is centered on the development and real-world evaluation of an explainable CDSS for arousal detection (see Section 2), requiring constant alignment of research priorities with evolving challenges and new insights and perspectives.

A major factor in discontinuing the prototype classification approach was the emergence of new and pressing challenges related to developing and optimizing an arousal detector based on practical clinical needs, as detailed in Section 6.4. Addressing these requirements, which culminated in the publication of ALPEC, took precedence over pursuing prototype-based arousal detectors.

Two further considerations made prototype-based explanations appear less appropriate for the objectives of this work.

First, arousal detection is a high-throughput task (in the patient records used in the User Study, a patient had around 300 to 500 arousals per night; see Table 1 in Appendix D) in which individual decisions carry less weight than for many tasks encountered in domains such as radiological diagnosis, and timeliness is essential due to patient waiting lists. This context motivates streamlined explanation methods that are expected to impose lower cognitive load compared to prototype-based explanations such as causal feature attributions techniques. As discussed in Section 5.1.2, practical experience reveals that clinicians are more likely to disregard explanations with high cognitive load.

Second, as examined in Section 4.3.2, post-hoc explainability for black box models has become the predominant practice, far surpassing ante-hoc explainability approaches. Also, despite the ongoing interest of research on prototype classification methods in the image domain, the analysis in Section 5.3.2 indicates that out of roughly 40 explainable

AI-enabled CDSSs in practical radiologic imaging, only one system made use of CBR. In a field where application-grounded user studies are still rare (see Section 1), selecting widely adopted methods enhances the relevance and generalizability of research outcomes. A comparison of both approaches was not feasible within the available resources, as will be further discussed in Section 7.4.4.

In summary, the decision was made to focus on causal feature-attribution methods rather than prototype-based explanations, and to employ a ML model with established performance for arousal detection (see Section 6.2).

The progression from SPARROW to later publications thus marks a shift in research priorities from methodological advancements in XAI to a focus on user-centered and domain-specific questions, mirroring a broader trend in the field of XAI (see Section 4.3.4). At the same time, this transition is reflected in the movement from functional-grounded evaluation in SPARROW to application-grounded evaluation in the User Study, with a continuous emphasis on evaluation methods providing a consistent thread throughout the sequence of publications.

7.2 Comprehensive Polysomnography (CPS) Dataset: A Resource for Sleep-Related Arousal Research

7.2.1 Contributions and Relevance of the Findings

The CPS publication is closely linked to the subsequent ALPEC publication. ALPEC is the first publication to utilize the CPS dataset and includes further analyses and information that are included in this section to inform the following discussion.

As detailed in the related work section of ALPEC (Section 2.3 in Appendix C), the CPS dataset represents a uniquely comprehensive resource for sleep-related arousal research, comprising 17 raw PSG channel modalities along with a broad set of derived features. It incorporates innovative physiological signals, including pulse transit time and beat-by-beat blood pressure estimates, which are rarely available in public PSG datasets and have the potential to advance sleep diagnostics [18, 104, 118]. The dataset further provides standardized questionnaire data, detailed annotations specifying whether arousals were first identified via EEG or other physiological changes, and a range of clinically relevant medical outcome variables. This depth and breadth support diverse research directions in ML, clinical, and translational studies, particularly for projects that seek to leverage new modalities or enhance understanding of arousal detection across varied patient groups.

To ensure the dataset is readily accessible, the PhysioNet page [75] offers several supporting tools. Croissant specifications are included, delivering a standardized metadata format that supplies a unified, ML-specific description of the dataset, thereby improving discoverability and usability across different tools and platforms [10]. Python scripts with installation and usage instructions are available to facilitate flexible access to patient data, covering channel measurements, events, and questionnaire data. Additionally, summary

statistics on questionnaire data can be efficiently generated using the SweetViz library [27].

Appendix H of ALPEC presents a detailed overview of dataset demographics and key domain-specific variables, alongside a discussion of representativeness. Appendix J provides a *Datasheet for Datasets*, a standardized documentation framework that enhances transparency, accountability, and reproducibility in ML by supplying structured descriptions of, e.g., the dataset's motivation, composition, collection process, and recommended uses. This framework supports informed decision making and helps reduce unintended bias [52].

Taken together, the CPS publication offers a uniquely comprehensive and accessible resource for sleep-related arousal research. By releasing high quality, richly annotated data, these efforts enable future studies that are aligned with the objectives of the underlying medical study [156] and serve to advance standards of transparency, reproducibility, and collaboration within the field. As discussed in Sections 1 and 5.2.3, the limited availability of public datasets, especially in underexplored domains, remains a major barrier to advancing solutions suitable for clinical practice. The availability of the dataset and associated tools lowers barriers for interdisciplinary research, supports robust benchmarking of new algorithms, and encourages exploration of innovative analyses and clinical applications. This collective progress benefits both researchers and practitioners by advancing understanding, detection, and treatment of sleep-related disorders, thereby contributing to progress in sleep medicine as outlined in the objectives of this doctoral work (see Section 2).

7.2.2 Link to Subsequent Publications

Within the broader scope of these doctoral studies, the collection of the CPS dataset is central. Its analysis motivates a key challenge on clinical annotation constraints that is addressed in ALPEC. Also, it serves as the primary resource for training arousal detection models in ALPEC and provides a realistic empirical foundation for the User Study. This ensures that professional sleep scorers, participating in the application-grounded user study, interact with data that closely reflects what they encounter in routine clinical practice. This approach strengthens the quality of the evidence of subsequent findings, as will be discussed further in Section 7.4.3.

7.3 ALPEC: A Comprehensive Evaluation Framework and Dataset for Machine Learning-Based Arousal Detection in Clinical Practice

To enhance the flow, the order of the next subsections is inverted relative to previous sections, as the general contributions build directly on the advances realized through ALPEC in preparation for the User Study.

7.3.1 Link to Subsequent Publications

The primary aim of this work was to develop a performant ML model trained on the CPS dataset for use in the subsequent User Study. Throughout this process, several previously

unrecognized challenges and gaps in the literature were identified and addressed. The resolution of these challenges is especially pertinent for the User Study and will be discussed in the following. For a detailed description of the challenges, refer to Section 6.4.

The first challenge addressed is the alignment of the arousal detection task with clinical annotation practices by shifting the focus from identifying complete events to localizing arousal onsets. This approach resolves a fundamental mismatch between the conventions of clinical annotations and standard practices in training ML models for arousal detection, thereby ensuring that the model development process is consistent with the realities represented in the CPS dataset. We show that using continuous segmentation with *DeepSleep 2.0* [40, 81] in combination with the proposed interval-based onset detection training scheme results in performance comparable to complete event detection, all while honoring practical annotation constraints.

The second and third challenges concern the absence of standardized performance evaluation and the insufficient ethical and operational scrutiny given to the task in related research. To address both issues, the *approximate localization and precise event count* (ALPEC) framework is introduced for post-processing and performance evaluation and to optimize arousal detection models for clinical requirements. The F2-score is advocated as the principal performance metric, aligning evaluation with ethical considerations and operational goals typical of CDSSs. Rather than supporting autonomous use, this evaluation foregrounds decision support, where a CDSS highlights candidate arousal events for clinician review while preserving diagnostic oversight. This approach meets ethical principles in medical AI and anticipates regulatory mandates for human supervision in healthcare applications (see Section 1 in Appendix C). This topic is further discussed in detail in Section 7.5. ALPEC is thoroughly compared to alternative evaluation approaches and is used to optimize the *DeepSleep* model trained on the CPS dataset.

A fourth challenge involves the mostly untapped potential of using many modalities in arousal detection research. This is addressed by leveraging the *DeepSleep* model architecture, which natively supports the integration of in principle any number of multimodal data channels [81]. Using the ALPEC framework, the model candidate utilizing the richest modality set showed significantly better performance compared to other candidates. The top-performing multimodal model was selected for use in the subsequent User Study.

In summary, this work establishes the methodological and empirical foundation for the subsequent User Study.

7.3.2 Contributions and Relevance of the Findings

From the perspective of these doctoral studies, the ALPEC framework was developed to facilitate the User Study, yet it also offers broader contributions.

In ALPEC, we advocate for a focus on detecting arousal onsets rather than complete events to achieve greater alignment with clinical annotation standards. Such alignment

enhances the practical relevance of model training and evaluation and improves interoperability across laboratories.

Moreover, our experimental results demonstrate that the additional modalities provided by the CPS dataset significantly enhance performance compared to models restricted to the most common channels in the literature. Furthermore, excluding electrode-based channels yields results comparable to those of a univariate EEG-based model. This finding suggests that reliance on electrode-based measurements, which are prone to artifacts and displacement, could be reduced and might facilitate home-based diagnostic practices (see Section 5 in Appendix C). This experiment represents a preliminary investigation into a central goal of the underlying medical study: the feasibility of employing alternative data modalities (see Section 6.3).

The ALPEC framework is introduced as a standardized performance evaluation and post-processing framework for tasks requiring approximate localization and precise event count detection. It enables more accurate performance assessment than conventional window-based or pointwise methods while avoiding common methodological pitfalls, such as class imbalance and the lack of cross-subject validation. Critically, ALPEC supports varied annotation formats, including point annotations, constructed intervals, and events with explicit bounds, making it relevant across a wide range of temporal event detection problems. Its tunable hyperparameters allow adaptation to task-specific requirements. For event detection tasks with different practical needs, this framework can serve as an adaptable blueprint, drawing on a relevant taxonomy for performance evaluation in time series data by Sørbø and Ruocco [140].

In conclusion and in reference to the objectives of these doctoral studies (see Section 2), this publication advances the field of sleep medicine by presenting a pathway to align arousal detection methods with clinical practice, demonstrating the performance benefits of multimodal approaches, and demonstrating the feasibility of moving beyond exclusive reliance on electrode-based modalities. Finally, ALPEC generally supports the development and optimization of AI-based event detection systems, that requires both approximate event localization and precise event counts, featuring adjustable hyperparameters, which makes it applicable to a wide variety of tasks.

7.4 Assessing the Real-World Utility of AI-Powered Decision Support for Arousal Diagnostics: An Application-Grounded User Study with Professional Sleep Scorers

7.4.1 Summary and Discussion of the Results

We provide a comprehensive evaluation of human-AI collaboration in clinical arousal scoring, examining the effectiveness of AI assistance, the influence of transparency (white-box versus black-box), and the impact of workflow integration according to the timing of AI input, whether offered from the outset or as a post-hoc quality control step.

In the first set of experiments, we evaluated human-AI collaboration performance with quality control support by pooling outcomes across both AI transparency conditions. Our findings indicate that the utility of AI support depends on both the evaluation standard and workflow context. When assessed against the CPS reference standard, which is the training basis for the ML model, human-AI teams achieve significantly higher performance and greater consistency than unaided human experts. This suggests that AI assistance can effectively align human decisions with established clinical standards. In contrast, evaluation against the more heterogeneous consensus reference reveals no measurable improvement, a result discussed further in Section 7.4.2.

The remainder of the results focus on assessments against the CPS standard. Across both principal evaluation approaches, event-level and total event count-based, the results exhibit similar patterns, though not all trends attain statistical significance. A detailed discussion can be found in the manuscript's comprehensive analysis (Section 5 in Appendix D). Here, we highlight the primary trends and principal findings.

The results demonstrate that transparent AI, especially when used as a quality control measure following an initial manual scoring, enhances both the collaborative performance of human-AI teams and the consistency of their results when compared to black-box support. User experience evaluations indicate that transparent AI support is consistently and substantially rated higher than black-box support across all assessed dimensions, including usefulness, confidence, trust, validation ease, and enjoyment. Transparent explanations emerge as essential for clinical acceptance, with most participants regarding the current system as nearly ready for practical use, pending minor refinements.

Although post-hoc timing of AI assistance produces the most accurate outcomes, three quarters of participants express a preference for having AI suggestions available from the outset, citing workflow efficiency and increased confidence. This subjective preference is corroborated by efficiency measurements, which identify a key challenge: both transparent AI support and quality control are generally associated with approximately double the median task duration. The range of task durations is wider for transparent AI support, suggesting potential for improvement, whereas sustained efficiency gains in the quality control configuration may be limited by the workflows applied in the study. A more detailed discussion of the implications of workflow design, along with potential future directions, is conduced in Section 7.5.

Together, these results highlight the value of transparent and well-timed AI assistance in clinical decision support. Explanations transform AI from a passive source of suggestions into an actionable partner that users perceive as more trustworthy, useful, and easier to validate. The study further underscores the importance of curating clinical standards and carefully designing workflows to achieve a balance between efficiency and accuracy. In summary, this work positions human-AI collaboration in arousal scoring as a complex design problem, where model development, usability, workflow, and governance must be addressed in concert to achieve effective clinical decision support.

Contributions. This work represents the principal contribution toward achieving the objectives outlined in Section 2. Building on the arousal detector developed within ALPEC, it offers a real-world evaluation that demonstrates the significant practical value of explainable AI-based CDSS for arousal detection, contributing to advances in sleep medical diagnostics. By conducting an application-grounded evaluation, the User Study addresses a critical need in the XAI field (see Section 1) and provides robust empirical insights to inform the future development of explainable decision support systems.

7.4.2 Comparison of Findings to Existing Literature

We now reflect on key design choices, contributions, and findings of the User Study in the context of the existing literature, as introduced in the opening sections of this dissertation.

Trust and Over-reliance on AI. As outlined in Section 5.1.1, high levels of trust in AI can foster over-reliance. Our results corroborate this, particularly in the case of one participant (*HU-5*) who reported an impression that the AI was correct in 95 percent of cases, yet showed the lowest collaboration performance in most scoring regimes. Across all participants, we observed a substantial increase in trust when moving from black-box to transparent support (from 7.2 to 8.8 Likert points). Importantly, this increased trust did not result in adverse outcomes for most participants. On the contrary, collaborative performance significantly improved with transparent AI support compared to black-box support, especially at the event-level evaluation. Furthermore, three quarters of participants reported being influenced by the transparent AI, suggesting active conscious engagement with the AI assistance rather than passive or unconscious influence. Nonetheless, I concur with one participant's caution that AI assistance offered from the outset may increase the risk of over-reliance, particularly for less experienced scorers, which underscores the advantage of using AI as a tool for quality control. The topic of workflow design is further examined in the final discussion in Section 7.5.

Section 5.1.2 discusses measures to address over-reliance, including providing explanations on demand and implementing cognitive forcing strategies. In our CDSS, participants could choose whether explanations appeared automatically for each event or only when triggered manually, via mouse click or keyboard shortcut. All participants preferred the manual option, which may have helped mitigate over-reliance. Cognitive forcing functions were not deployed in this study but could serve as a valuable addition in future investigations.

Alignment of Mental Models of Humans and AI. Section 3.3.2 emphasizes that explanations can support an alignment between the mental models of users and AI, thereby fostering trust and acceptance. While our study did not evaluate alignment at a detailed feature level, we found that transparent AI broadly increased alignment between human scoring behavior and the standard represented by the ML model. All participants found

the explanations fully understandable and helpful, suggesting that a meaningful alignment of mental models took place. Additionally, explanations enhanced user acceptance of the AI model. This is further illustrated by one participant, who was surprised that the AI did not place greater emphasis on increases in EEG frequency, in contrast to human scorers following clinical guidelines [26]. This is indicative of *counterintuitive explanations*, which is a known failure mode for explanations as discussed in Section 4.3.3. However, this did not undermine trust, as participants were able to adapt, and the overall perceived plausibility of the explanations remained high (8.5 out of 10 Likert points).

The subjective assessment of transparent AI versus black-box AI support, 8.9 against 7.4 Likert points, aligns with the objectively measured improvement in collaboration performance. Accordingly, we did not identify a notable perception-behavior gap, as discussed in Section 4.4.2.

Inter-rater Variability and Performance on Different Datasets. As discussed in Section 3.4, high inter-rater variability and reduced model performance on new datasets are persistent challenges in arousal scoring. Both phenomena were evident in our study. We observed substantial inter-rater variability among participants and pronounced differences in scoring patterns relative to the CPS reference standard. This variability was identified as the primary reason for low performance by both the AI model, trained on the CPS standard, and the human-AI teams when evaluated against consensus annotations. As a result, an AI that was unmatched by human scorers on the CPS standard performed below human level when benchmarked against the consensus.

Interpretability-Completeness Tradeoff. Section 4.3.2 addresses the well-known tradeoff between interpretability and completeness in explanation design. As noted there, the literature suggests providing and evaluating explanations at different levels of detail. We implemented this by offering three levels of explanation detail, which modulate the attribution value threshold above which feature attributions are displayed. This allowed users to select their preferred level of detail. While limited statistical power prevented formal evaluation along this tradeoff, user preference data were collected. As observed in the User Study, most participants preferred the medium level of detail, with two opting for the highest level.

7.4.3 Validity and Scope and of the Evaluation Approach

We assert that our user study constitutes an application-grounded evaluation which is considered the most direct and contextually relevant method available (see Section 4.4.3) and in the XAI community is often assumed to be the gold standard.

In the following, we critically appraise these assumptions and outline further options for evaluation.

Application-Grounded Evaluation. Although we describe the study as application-grounded, several simplifications were necessary as detailed in the User Study manuscript. To save time, experts assessed only the initial segment of each measurement, approximately three hours of data which is the period where arousals are typically most frequent, and the task was limited to binary arousal detection. In standard clinical practice, however, scorers not only detect arousals but also annotate their causes, such as leg movements, respiratory events, or snoring, and document additional findings including sleep stages and various respiratory occurrences like apnea or hypopnea. For the participants in this user study, such additional annotations and events were made available in a pre-scored format, again, to save time. Importantly, as stated by Doshi-Velez and Kim [36] who introduced this taxonomy of evaluation approaches, a task that is simplified or shortened can still qualify as application-grounded, and we assert that our approach aligns with this perspective.

In addition, the CDSS was developed with guidance and oversight from an expert in sleep medical diagnostics to ensure it included all features necessary for expert scoring. The system supports extensive customization, as is typically required by arousal scorers, and was designed for an intuitive, user-friendly experience, presenting all relevant information in a single view most of the time and including keyboard shortcuts for efficiency. The high ratings for usability and user experience in the User Study support both the usability goal and the classification of this work as application-grounded.

Quality of the Evidence. Beyond application-grounded evaluation, it is important to consider whether alternative approaches could further strengthen the evidence for the clinical utility of the CDSS.

The hierarchy of evidence for AI and XAI empirical studies proposed by Famiglini et al. [39] defines ten levels, with level 1 as the highest quality evidence and level 10 the lowest. Our study is situated at level 3 which they define as "Single quasi-experimental study (e.g., nonrandomized, with concurrent or historical controls) involving prospective real-world cases considered by real practitioners in real-world settings." [39].

We implemented control conditions addressing both transparency and timing by prompting practitioners to score with either black-box or transparent AI support, both from the outset and as a post-hoc quality control step. Prospective studies, in contrast to retrospective approaches, collect data specifically for the purposes of the study rather than relying only on preexisting data, as was the case in our design. Also, as stated previously, our study aligns with the criteria for application-grounded evaluation, thereby meeting the requirements of real-world cases, real practitioners, and a real-world setting.

The evidence provided by our outcomes could be improved in two major ways. First, conducting a randomized controlled trial (RCT, level 2) instead of a non-randomized study could strengthen validity. Second, a meta-analysis of RCTs in prospective, application-grounded settings (level 1) would provide the highest quality evidence.

A RCT would require presenting a randomly assigned subset of participants with

the control condition, necessitating a larger sample size to preserve statistical power. Limitations in resources and feasibility prevented us from adopting this design (we discuss this further in Section 7.4.4).

A robust meta-analysis of RCTs is currently not feasible, as discussed in Section 1, due to the scarcity of relevant application-grounded user studies in the literature. Nevertheless, in Section 5.1, we undertook a limited analysis of existing work to identify both divergent findings and shared patterns. Also, the discussion of the User Study in the manuscript and this dissertation places our methodology and findings in the context of current literature and contributes toward enabling future meta-analyses.

Placebo Control. A further option to strengthen the evidence, that is not explicitly caputred in the hierarchy of evidence by Famiglini et al. [39], would be to include a placebo control. For instance, following the approach of Amarasinghe et al. [13], we might have included conditions that exposed participants to randomly selected features or irrelevant information in addition to the real post-hoc feature-attribution methods. The absence of such controls represents a limitation of the present study.

System Causability Scale (SCS). Another option would have been to incorporate the System Causability Scale (SCS) instrument into the questionnaires. As discussed in Section 4.4.3, the use of the SCS would have provided two main benefits. Its standardized structure would improve the comparability of our results, and it explicitly probes whether explanations provide causal insights, which our own questions did not address. This reflection is aimed at increasing the awareness of this evaluation method and encouraging its use in future studies on the benefits of XAI.

7.4.4 Reflections on Limitations

Key limitations of this study, as detailed in the User Study manuscript, are the small sample size, potential bias in the ground truth labels, limited generalizability of the CPS reference standard, use of a single model optimization strategy, reliance on a single explanation modality, possible residual learning effects, short user exposure duration, and the influence of specific user interface design choices.

This section aims to contextualize these limitations and considers additional factors regarding the scope of the User Study.

Given the available resources, it was necessary to design the study so that participation could be completed within approximately five to six hours per participant. This constraint fundamentally shaped several of the study's limitations, including the reliance on a single model optimization strategy, use of only one explanation modality, and brevity of user exposure. The following discussion expands upon these constraints.

As noted by Gunning et al. [63], it is preferable to introduce explanation modalities incrementally to discern the individual impact of each. For example, the effects of providing

a detailed confidence channel, which indicates confidence for each time point, or a high-level visualization of confidence for an arousal event, might differ in terms of scorer performance. Even finer distinctions, such as the difference between channel-wise importance, reflecting the rank order of channels by importance, and feature-wise importance for each channel, could have been assessed. One participant, for instance, observed that in situations where analytic area space on the CDSS interface was limited, additional channels could conveniently be accessed through the explanations, especially when classified as relevant by the AI. This feature was noted as particularly helpful.

Furthermore, only one ML model optimized for the F2-score was employed. Exploring alternative optimization objectives or models with different feature sets, such as the *D3* model from ALPEC that excludes electrode-based modalities, would have offered informative comparisons.

Finally, the study utilized only one feature attribution method. Investigating alternative explanation approaches, as discussed in Section 4.1.4 like prototype-based explanations, or directly comparing methods during the study, such as contrasting *DeepLIFT* with other techniques, could have yielded additional insights. Instead, only an exploratory post-hoc comparison with *GradientSHAP* was performed, as detailed in Appendix B of the User Study manuscript.

7.5 AI-Based Decision Support versus Autonomous AI Decision Making and Future Avenues

In an exchange with a clinician in sleep medicine regarding the User Study, I received the following statement: “I am convinced that in the medium term, AI will not only assist but also efficiently perform almost all of the scoring, which you kindly do not yet see.”. So, there is an elephant in the room that we have not yet addressed: the relevance of AI-based decision support as compared to autonomous AI decisions.

This topic, especially with respect to the evolving role of human experts, is expansive and exceeds the scope of a single section. Therefore, the following analysis hast to be selective rather than comprehensive. To provide a meaningful contribution, I will focus primarily on arousal diagnostics, draw on insights from the User Study, and consider the role of XAI within this context.

This section also intentionally adopts a more opinionated and narrative tone to elaborate on perspectives that I developed through my doctoral studies and in particular the User Study. By linking future directions for XAI to the broader discourse on AI-based decision support versus autonomy, my aim is to shed light on objectives for the field that are often underemphasized.

Given this focus, several central questions arise:

1. Should workflows continue to prioritize human expertise, potentially supported by AI, or should humans be removed as central actors in arousal scoring and be replaced

by fully autonomous decision making by AI systems?

2. If continued human involvement is warranted, what would be the roles and tasks of humans, and how should future workflows be designed?

3. What are the implications for the future role of XAI?

4. And how might this discussion shape the future of research on human-AI collaboration in arousal diagnostics?

These questions are considered in the following sections, following the order outlined above.

7.5.1 Investigation of Autonomous AI Decision Making and the Role of Human Involvement.

This section examines arguments both supporting and opposing autonomous decision making by AI, and considers the significance of human involvement in high-risk tasks.

Reasons for Autonomous AI Decisions. The primary arguments in favor of autonomous AI decision making center on anticipated improvements in result quality and time efficiency, both of which are intended to benefit patients. Other factors, such as the shortage of trained personnel, are acknowledged but will not be discussed in detail here.

Let us first consider the aspect of **quality**, drawing on findings from the User Study.

Results from the User Study demonstrate that at the event-level, which is most relevant to human scorers, the AI alone significantly outperformed all human-AI teams (see Table 7 in Appendix D) on the CPS reference standard. Interpreting this result requires caution, as the CPS reference is not a curated widely accepted standard but a composite from practice, reflecting the approaches of multiple scorers. High inter-rater variability in arousal scoring in general (see Section 3.4) and our own assessments (e.g. Figure 8 in Appendix D where the CPS reference is included as *annotator 9*) suggest that results could differ with a more carefully curated standard aligning more closely with typical human scoring behaviors.

Nevertheless, Buçinca et al. [30] point to mounting evidence that human-AI teams frequently underperform compared to AIs acting alone. Another argument in favor of more AI autonomy in arousal scoring can be drawn by analogy to radiology which already features at least one FDA-cleared autonomous AI-based computer-aided detection system in practice (see Section 5.3.2). Also, while arousal detection is a high-risk task as arousals are linked to severe health conditions (see Section 3.4), it differs from many radiological tasks in that it is also a high-throughput task and each decision individually carries less weight (patients often have hundreds of arousals per night, see e.g. Table 1 in the User Study in Appendix D). Thus, by analogy, a higher level of automation may be acceptable for arousal detection than in tasks like lesion identification on radiological images. Together with the findings from radiology, where AI systems have reached superhuman accuracy

(see Section 1), the issue is likely no longer whether AI can outperform humans in arousal detection, but rather when this becomes standard, if not already the case.

Now, let us turn to the **time efficiency** aspect, again informed by findings from the User Study, which offers several insights regarding the efficiency of human-AI collaboration.

As discussed in Section 7.4.1, the quality control workflow, in which human solo scoring is followed by AI-assisted review, combined with transparent AI assistance, yielded a significant improvement in event-level result quality and a substantial reduction in the spread of results compared to black-box assistance. Notably, because human raters completed initial scoring without any AI input, this workflow may also mitigate risks associated with over-reliance (see Section 7.4.2). Despite these benefits, the quality control approach imposed significantly greater time demands, and participants typically preferred assistance from the outset (see Section 7.4.1).

To address time and acceptance challenges, one possible strategy would be for the CDSS to provide an easily navigable difference view between human-scored and AI-suggested events. This could improve efficiency and heighten the practical value of AI support, thereby increasing user acceptance.

Still, the fact remains that this workflow inevitably increases the amount of time required relative to human solo scoring.

As many in the healthcare sector can attest: when asked how accurate an AI-based diagnostic system needs to be to be considered acceptable in practice, physicians often say 100 percent or close to it. While this view is understandable from the perspective of a clinician responsible for patient care, it is unhelpful. Although physicians may appreciate outcomes as illustrated above, a system that achieves higher result quality yet requires more time for scoring than manual methods would likely be considered suboptimal by healthcare providers or health insurance companies, particularly as more autonomous AI systems are introduced.

To this point, the evidence suggests that autonomous AI systems confer advantages in both result quality and time efficiency for arousal scoring over human-AI teams. However, several important issues concerning autonomous AI systems also warrant careful consideration, which we address next.

Issues with Autonomous AI Decision Making and the Importance of Human Involvement. AI systems in productive use, face several significant challenges, which are expecially problematic in high-risk environments:

- AI systems may encounter out-of-distribution cases [14, 126], for example, due to artifacts in data collection such as electrode displacement in PSGs recordings [70], or because of patient characteristics not represented in the training data.
- Model drift can lead to decreasing predictive quality, for example as a result of data drift due to new sensors or protocols, or due to model updates, as is possible

with Predetermined Change Control Plans (see Section 4.5.1), which could lead to degraded performance.

- Models can learn shortcuts, exploiting spurious correlations rather than meaningful patterns [53], a documented problem in practice where, for example, some radiological models focused on irrelevant covariates [91, 128].
- Predictions may be biased, for example due to biased training data [110, 131].
- AI systems can be subject to adversarial attacks [144].
- System outages or technical failures may render AI unavailable.

Since regulatory requirements for human oversight and monitoring remain broad and depend on specific risks and clinical contexts (see for example Section 4.5.3), it could be argued that relying solely on automated safeguards, alongside indirect supervision where humans are alerted only if a problem is detected, would be sufficient to mitigate these issues.

However, in my view, automated strategies should complement human oversight rather than replace it, and direct human involvement in core tasks remains essential.

Several considerations support this stance. As noted above, AI systems can become unavailable, making it necessary for humans to take over. More fundamentally, removing humans from processes such as arousal scoring would, over time, erode the expertise needed to conduct the task manually during outages or unforeseen challenges. When any of the listed issues arise and cannot be quickly resolved, the absence of adequately trained professionals would jeopardize patient care.

Adapting to evolving medical knowledge and shifting guidelines also depends on experts who can curate and update training data. Although a dedicated specialist team could in principle manage updates separately from routine operations, this approach would introduce unavoidable delays in implementing new diagnostic guidelines, as it would require new models to be developed, validated, potentially re-certified, and deployed. This process can be expected to be slower than retraining clinical staff in updated protocols, ultimately delaying the translation of scientific advances into routine care and undermining patient benefit. Section 5.2 highlights challenges such as skill degradation among practitioners and the need to maintain up to date knowledge bases. Both are well recognized barriers to the adoption of AI-enabled CDSSs, and these challenges are likely to be even more significant in the context of autonomous decisions from AI systems.

Moreover, it is clear that automated solutions can not detect all problems, especially as novel or unanticipated issues arise. In such contexts, humans trained and equipped with the appropriate tools (as will be discussed in Section 7.5.3) should however be able to adapt and respond effectively. Employing one AI system to monitor another may offer research opportunities and contribute to risk mitigation. However, such approaches do

not replace the need for human oversight, as they merely shift the requirement for human involvement to a different layer and perpetuate the risks associated with removing humans from the core task.

In summary, the risks associated with a strong dependence on AI systems underscore the importance of sustained human involvement on the actual tasks. Accordingly, the following discussion addresses how to design workflows that may effectively integrate human expertise, specifically in the context of arousal scoring.

7.5.2 Towards Efficient Human-AI Collaboration in Arousal Scoring

Thus far, the evidence indicates that AI systems alone are likely to outperform humans in arousal scoring. However, several critical considerations necessitate ongoing human involvement in this task. This observation raises the question of how these insights can inform the design of workflows for human-AI collaboration that are both efficient and safe.

The key idea is to grant greater autonomy to AI systems, with findings from the User Study illustrating a possible implementation of this approach. In addition to the quality control workflow discussed in Section 7.5.1, the User Study also investigates a workflow where AI assistance is provided from the outset. Although this approach on its own did not yield significant performance improvements, regardless of the transparency of assistance (see the *Start* regime in Table 11 in Appendix D), it paves the way toward workflows approaching increased AI decision autonomy.

It is important to clarify that in the User Study, participants were given neither instructions to evaluate *every* 30-second interval (the traditional manual approach), nor explicit permission to *bypass* intervals. They were simply prompted to use AI assistance and shown how to quickly navigate between AI-suggested events. In practice, seven of eight participants defaulted to reviewing each interval when scoring with AI assistance from the outset, only using interval-jumping in the quality control stage. Remarkably, one scorer (*HU-5*) shifted strategy in the final scoring pass (using black-box AI assistance from the outset), opting to jump directly between AI-suggested events. This allowed scoring of three hours of data in about five minutes (well below one second per interval on average), essentially using a quality control approach from the start. Actually, by optimizing the ML model for the F2-score in ALPEC (see Section 7.3.1) we intended to encourage this very behavior.

The emergence of this behavior is therfore notable. Although *HU-5* developed over-reliance, accepting nearly all AI-suggested events, in that scoring pass, this scorer achieved the best performance of all human-AI teams and even surpassed the solo AI (see Figure 16A in Appendix D, where *HU-5* appears as the upper whisker in the *Start-BB* boxplot and elsewhere as an outlier; note also that the AI itself performed strongly on this patient record, which contributed to this outcome).

Next, it is instructive to examine more closely how this proposed workflow affects time

efficiency.

We can make a rough estimate based on Table 6 in Appendix D, combining two patient records used for quality control and covering approximately six hours of recordings. Typically, with 30-second intervals, scorers would have to manually assess 720 intervals. The AI model in this case marked 130 true positives, 80 false positives, and 30 false negatives. If scorers were to conduct quality control only for AI-suggested events, they would need to review just 210 intervals. Ignoring cases where an AI-predicted event spans more than one interval and assuming scoring time is constant per interval, regardless of interval content or AI assistance, this workflow could reduce time spent on scoring by roughly 70 percent compared to the current manual approach which is a substantial gain.

To summarize, a collaborative workflow exploiting partly autonomous AI decisions along these lines holds the potential to achieve a good team performance with substantially decreased human time investment as compared to unaided human scoring. Now the question is how the issues identified in Section 7.5.1 can be adressed for this workflow to be efficient *and* safe. This leads us to investigating the role of XAI in this workflow more closely, leading to future directions for the field of XAI in Section 7.5.3 and human-AI collaboration in arousal scoring in Section 7.5.4.

7.5.3 Future Directions for the Field of XAI.

Let us start by summarizing the current state of the field of XAI. On a broad level, and informed by the examinations in this dissertation, the field of XAI currently faces significant challenges. One might say that it faces something resembling an identity crisis. Foundational concepts and terminology remain contested (Section 3.2), and the question of what constitutes an explanation, or a good explanation, remains notoriously challenging (Section 4.4.1). Ongoing debates, such as the merits of ante-hoc versus post-hoc explanation and the interpretability-accuracy tradeoff (Section 4.3.2), continue to divide opinion. Evidence for what works in practice is mixed, and reliance on proxy measures or tasks clouds the picture further (Section 5.1.1). Regulatory frameworks do not mandate specific explainability approaches (Section 4.5.6), and for good reason, as explanation design must remain sensitive to context (Section 4.3.4). Meanwhile, high-risk AI systems are entering practice with minimal attention to explainability (Section 5.3.2), prompting questions about the field's future direction.

Given that the intrinsic value of an explanation remains difficult to define, human-grounded evaluation of XAI approaches typically focuses on the effect of explanations on collaborative performance. This encompasses task and domain-specific measures, complemented by subjective criteria such as user satisfaction, trust, and acceptance (see Section 4.4). A risk that comes with this, however, is the tendency to conflate the value of explanations solely with gains in task accuracy.

Nevertheless, as established in the literature, XAI can facilitate the identification of

all issues associated with AI systems described in Section 7.5.1 (where the detection of system outages is obviously trivial), which is achieved by exposing salient features and model reasoning [97, 123, 126, 128].

As I argued in Section 7.5.1, maintaining direct human involvement in core tasks like arousal scoring remains important.

Consequently, it appears reasonable to suggest that the primary value of explanations may lie in enabling clinicians to recognize problematic behaviors of AI systems. This perspective aligns with the expectations of regulatory bodies, which, while not mandating specific explainability approaches, emphasize functional transparency for objectives such as bias prevention, fairness, oversight, safety, security, accountability, and monitoring (see Section 4.5). XAI can advance these regulatory aims.

Adopting this approach is not without obstacles. Industry may resist practices that reveal flaws, as such findings entail costs related to system modifications or even withdrawal from service. As Rudin [128] highlights in a well-known critique, medicine has seen a trend toward widespread, not always critical, adoption of black box models, sometimes driven by commercial incentives, despite potential risks such as failures that arise from training data deficiencies. To clarify, I do not advocate exclusively for intrinsically interpretable models in detecting such issues. Where post-hoc explanations are used, caution is needed regarding risks like fairwashing, where post-hoc explanations are manipulated to obscure ongoing unfair decisions [9, 22]. A focus on identifying dangerous or problematic AI behaviors could actually guide efforts to empirically compare the utility of ante-hoc and post-hoc approaches. Building benchmark datasets that reflect real-world challenges and embedding these evaluations within application-grounded tasks, as will be outlined for arousal detection in Section 7.5.4, is a critical direction for future research.

Thus, the existing approach to evaluating explanations remains appropriate and does not require fundamental revision. Rather, the definition of business goals, or in clinical contexts the clinical objectives, should be broadened to explicitly encompass the identification of unsafe and problematic AI behavior.

Addressing the anticipated industry reluctance, it is important to note that regulatory bodies, while not mandating specific explainability techniques, nonetheless expect risk-mitigating best practices. By empirically demonstrating in application-grounded studies that particular explainability approaches offer practical value for functional transparency, their adoption could become a de facto requirement for certification, even absent explicit regulatory mandates.

It is also essential to acknowledge that XAI should not be viewed as a panacea. For example, issues such as adversarial attacks can deceive post-hoc explanation methods [22, 139], so explainability must complement, not replace, other safeguards such as clear communication of system limitations and meticulous record-keeping, as required for example by the AI Act (see Section 4.5.3).

In summary, focusing on whether and how XAI can support the safe adoption of

increasingly autonomously deciding AI systems in high-risk settings may offer a renewed sense of focus for the field. This entails maintaining meaningful human involvement in critical diagnostic workflows, while drawing on the predictive power of AI systems.

7.5.4 Future Directions for Human-AI Collaboration in Arousal Scoring.

Building on the previous sections, advancing human-AI collaboration in arousal diagnostics could be achieved through a follow-up study with several key characteristics as outlined in the following.

An AI model, trained on a large and well-curated dataset, optimized for relevant task requirements using ALPEC, and also towards minimizing false negatives, provides arousal predictions. Human experts are then assigned solely to correct false positives among AI-suggested events, a design that has the potential to enhance collaborative performance while being time-efficient (see Section 7.5.2). In this context, meeting a specified performance threshold in human-AI collaboration would serve as a quality gate rather than the primary objective of the study. Instead, synthetically generated issues, such as out-of-distribution cases, adversarial samples, or biased predictions, are intentionally included in the dataset (see Section 7.5.1). It can thus be determined whether clinicians, with or without explanations of AI predictions, can identify these issues, for instance by being mandated to diverge from the quality control workflow and reverting to manual scoring upon detecting any problems with the AI predictions. Mitigating over-reliance should be addressed as well by utilizing mechanisms such as cognitive forcing functions (see Section 5.1.2). Suitable controls would include established functional transparency approaches resembling the current practice (see Section 4.5). Conducting such a study would require meticulous study design, careful selection of issues to include, thoughtful choice of XAI approaches, and comprehensive participant training in the use of the system. Positive outcomes in this study, would provide strong evidence for the value of XAI tools and human involvement in ensuring safe AI-based arousal scoring.

As discussed in Section 7.5.2, such a workflow would involve relatively little human time investment compared to manual scoring while preserving human expertise and accountability and keeping the dependency on AI systems at an acceptable level.

When paired with a high-performing AI system, this workflow and division of responsibilities promises to establish a favorable balance between conventional human-only scoring and fully autonomous AI decisions without direct human oversight, fostering the development of effective, efficient, and safe human-AI collaboration in high-risk settings.

8 Conclusion

This dissertation set out to advance the understanding and practical application of explainable artificial intelligence (XAI) and clinical decision support systems (CDSSs), with a particular emphasis on arousal diagnostics and its evaluation in real-world contexts. Ini-

tially, it established theoretical and practical foundations of these fields that contextualize the subsequent research and discussion. The four included publications collectively pursue this overarching aim, emphasizing evaluation approaches and alignment with the specific needs of arousal scoring in clincal practice.

The SPARROW study [74] introduces a methodology that enhances the semantic coherence of prototype-based classifiers and enables functionally-grounded assessment, thereby expanding the methodological repertoire for XAI. The CPS publication [75] delivers a comprehensive and well-documented polysomnography resource, supporting transparent and reproducible research on arousal detection and related clinical challenges, and serving as a foundation for the subsequent investigations. The ALPEC [77] work addresses core challenges arising from clinical requirements and introduces a post-processing and performance evaluation framework for temporal event detection, optimizing and assessing the performance of ML models intended for clinical decision support in arousal detection. Finally, the clinical User Study [76] provides empirical, application-grounded evidence on the utility of integrating explainability for AI-assisted arousal scoring. It demonstrates that transparent assistance, especially when provided as post-hoc quality control, improves human scoring performance and consistency under the CPS standard, while also revealing trade-offs in efficiency and user preferences.

Further discussion within this dissertation contextualized these findings in relation to the existing literature and offered additional perspectives on the study's scope and limitations. The following broader discussion considered the practical viability of AI-assisted arousal scoring workflows as autonomous AI decision making emerges, emphasizing the ongoing necessity of direct human supervision. It also highlighted future directions for both XAI and AI-assisted arousal scoring.

I hope that this dissertation will prove valuable for guiding future research and will support the development of meaningful and sustainable human-AI collaboration workflows in arousal diagnostics and related fields.

Glossary

Arousal Arousal refers to brief episodes of biological activation [122]. With respect to sleep, arousals are characterized by abrupt shifts in EEG frequency that last for at least three seconds and follow a minimum of ten seconds of continuous sleep [26]. iv, vi, viii, x, xi, 2, 3, 14, 15, 43–57, 59–63, 66–68, 110, 146

Artificial Intelligence "Artificial Intelligence systems can perform tasks that normally require human intelligence [...]. They can solve complex problems, learn from large amounts of data, make autonomous decisions, and understand and respond to challenging prompts using complex algorithms" [97]. 68, 70

Black box model "A black box model could be either (i) a function that is too complicated for any human to comprehend, or (ii) a function that is proprietary" [128]. 7, 16, 23, 24, 31, 49, 65

Comprehensibility With respect to a machine learning model: "[...] [C]omprehensibility refers to the ability of a learning algorithm to represent its learned knowledge in a human-understandable fashion" [19]. 6–9, 28

Comprehensive Polysomnography This term refers to the publication *Comprehensive Polysomnography (CPS) Dataset: A Resource for Sleep-Related Arousal Research* [75]. 70

Explainability With respect to a machine learning model: "Explainability is associated with the notion of explanation as an interface between humans and a decision maker that is, at the same time, both an accurate proxy of the decision maker and comprehensible to humans" [19]. viii, ix, 2–6, 8–10, 12–16, 21, 22, 29–34, 36, 39, 40, 42, 49, 64, 65, 67

Explainable Artificial Intelligence "Given an audience, an explainable Artificial Intelligence is [an artificial intelligence] that produces details or reasons to make its functioning clear or easy to understand" [19]. 72

Intelligibility Synonym for understandability [19]. 4, 6, 7, 11, 12, 33, 34, 69

Interpretability With respect to a machine learning model: "Interpretability [...] is [...] the ability to explain or to provide the meaning in understandable terms to a human" [19]. 6–8, 11, 12, 14–17, 22–27, 31, 36, 39, 43, 44, 49, 56, 64

Machine Learning Machine Learning is a subset of artificial intelligence and a "[...] branch of computational algorithms that are designed to emulate human intelligence by learning from the surrounding environment." [38]. 68, 69, 71

Model With respect to machine learning, a model is a computational system used to make decisions and predictions from input data. [97]. iv, viii, 1, 4–13, 15–18, 21–25, 27, 30–36, 44–46, 50–54, 56, 58, 59, 61–65, 67–69

Obstructive Sleep Apnea Obstructive Sleep Apnea (OSA) is a sleep disorder in which the upper airway becomes partially or completely blocked repeatedly during sleep, leading to interrupted breathing and reduced oxygen levels in the blood [44]. 71

Polysomnography "Polysomnography is the simultaneous recording of numerous physiological signals during attempted sleep, including activity of the brain, heart, eyes, and muscles. Polysomnography is considered the gold standard for the objective assessment of sleep and diagnosis of many clinical sleep disorders." [54]. 71

Transparency With respect to a machine learning model: "A model is considered to be transparent if by itself it is understandable" [19]. iv, 4, 6–9, 11–13, 20, 22–25, 31, 36, 40, 46, 51, 53, 54, 57, 63

Understandability With respect to a machine learning model: "Understandability (or equivalently, intelligibility) denotes the characteristic of a model to make a human understand its function – how the model works – without any need for explaining its internal structure or the algorithmic means by which the model processes data internally" [19]. 6, 7, 22, 25, 68

User Study When capitalized, refers to the publication *Assessing the Real-World Utility of Explainable AI for Arousal Diagnostics: An Application-Grounded User Study* (see Appendix D). 47–52, 55–61, 63, 67

Acronyms

AASM American Association of Sleep Medicine. 14

AI artificial intelligence. iv, viii–x, 1–9, 11–15, 18, 20–22, 27–31, 33–40, 42, 44, 46, 47, 50, 52–57, 59–67

ALPEC This term refers either to the publication *ALPEC: A Comprehensive Evaluation Framework and Dataset for Machine Learning-Based Arousal Detection in Clinical Practice* (see Appendix C) or to the corresponding methodological approach: *Approximate Localization and Precise Event Count*. 46–53, 55, 59, 63, 66, 67

CAD computer-aided detection. 42, 60

CBR case-based reasoning. 15, 17, 18, 48, 50

CDS Clinical Decision Support. 30

CDSS clinical decision support system. iv, viii, ix, 1–3, 5, 6, 9–11, 14, 15, 18, 21, 24, 26, 29, 30, 33, 34, 37–42, 46, 48–50, 52, 55, 57, 59, 61, 62, 66

CFE counterfactual explanation. 17, 18

CNN convolutional neural network. 45, 48

CPS Comprehensive Polysomnography. x, 14, 44–48, 50–54, 56, 58, 60, 67, 68

DARPA Defense Advanced Research Projects Agency. 4, 26

DeepLIFT Deep Learning Important FeaTures. 16, 17, 46, 59

DL deep learning. 11, 30

DNN deep neural network. 7, 48

DPIA Data Protection Impact Assessment. 30

DSI Decision Support Intervention. ix, 33

DSS decision support system. 3, 9, 15, 18–21

ECG electrocardiography. 14

EEG electroencephalogram. 14, 45, 49, 50, 53, 56, 68

EHR Electronic Health Record. 33, 38, 40, 41

EMG electromyography. 14, 45

EOG electrooculography. 14, 45

FDA Food and Drug Administration. ix, 29–34, 40, 42, 60

FHIR Fast Healthcare Interoperability Resources. 5, 39–42

GDPR General Data Protection Regulation. ix, 4, 12, 29, 30

GMLP Good Machine Learning Practice. 32, 34

HHS Health and Human Services. ix, 29, 33

HL7 Health Level 7. 5, 39, 41

IMDRF International Medical Device Regulators Forum. 32, 34

IRM Intervention Risk Management. 33, 34

LIME Local Interpretable Model-Agnostic Explanations. 4, 16, 17, 24

MDR Medical Device Regulation. 29, 30

MDSW Medical Device Software. 30

MHRA Medicines and Healthcare Products Regulatory Agency. 31, 32

ML machine learning. iv, ix, 4, 10, 11, 13–15, 17, 19, 20, 22, 25, 30–33, 50–52, 54, 55, 59, 63, 67

MLMD Machine Learning-Enabled Medical Device. 32

ONC Office of the National Coordinator for Health Information Technology. 33, 41, 42

OSA Obstructive Sleep Apnea. 14

PCCP Predetermined Change Control Plan. 30, 62

PPG photoplethysmography. 14, 45

PSG polysomnography. iv, 2, 14, 49, 50, 61, 67

PWA pulse wave analysis. 44, 45

RCT randomized controlled trial. 57, 58

SaMD Software as a Medical Device. 30

SCS System Causability Scale. 27, 29, 58

SHAP SHapley Additive exPlanations. 4, 16, 17, 24

SMART Substitutable Medical Applications and Reusable Technologies. 41, 42

SPARROW This term refers either to the publication *SPARROW: Semantically Coherent Prototypes for Image Classification* (see Appendix A) or to the corresponding methodological approach: *Semantically Coherent Prototypes for Image Classification*. x, 47–50, 67

SUS System Usability Scale. 29

XAI explainable artificial intelligence. iv, viii–x, 1–7, 10–17, 21–26, 29–32, 34, 36, 38, 39, 43, 44, 47, 48, 50, 55–60, 64–67

Bibliography

[1] Regulation (EU) 2017/745 of the European Parliament and of the Council of 5 April 2017 on medical devices, amending Directive 2001/83/EC, Regulation (EC) No 178/2002 and Regulation (EC) No 1223/2009 and repealing Council Directives 90/385/EEC and 93/42/EEC, 2017. URL https://www.medical-device-regulation.eu/wp-content/uploads/2019/05/CELEX_32017R0745_EN_TXT.pdf. Accessed: 2025-10-19.

[2] Art. 22 GDPR: Automated individual decision-making, including profiling, 2018. URL https://gdpr-info.eu/art-22-gdpr/. Accessed: 2025-10-19.

[3] Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act), 2024. URL https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng. Accessed: 2025-10-19.

[4] Epic Systems Corporation Products, 2024. URL https://showroom.epic.com/stage?id=35. Accessed: 2025-10-21.

[5] Health Data, Technology, and Interoperability: Certification Program Updates, Algorithm Transparency, and Information Sharing Rule (HTI-1), 2024. URL https://www.federalregister.gov/documents/2024/01/09/2023-28857/health-data-technology-and-interoperability-certification-program-updates-algorithm-transparency-and#p-723. Accessed: 2025-10-19.

[6] Michael D Abràmoff, Philip T Lavin, Michele Birch, Nilay Shah, and James C Folk. Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices. *NPJ digital medicine*, 1(1):39, 2018.

[7] Amina Adadi and Mohammed Berrada. Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). *IEEE access*, 6:52138–52160, 2018.

[8] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. *Advances in neural information processing systems*, 31, 2018.

[9] Ulrich Aïvodji, Hiromi Arai, Sébastien Gambs, and Satoshi Hara. Characterizing the risk of fairwashing. *Advances in Neural Information Processing Systems*, 34:14822–14834, 2021.

[10] Mubashara Akhtar, Omar Benjelloun, Costanza Conforti, Luca Foschini, Joan Giner-Miguelez, Pieter Gijsbers, Sujata Goswami, Nitisha Jain, Michalis Karamousadakis, Michael Kuchnik, et al. Croissant: A metadata format for ml-ready datasets. *Advances in Neural Information Processing Systems*, 37:82133–82148, 2024.

[11] Ahmed Shihab Albahri, Ali M Duhaim, Mohammed A Fadhel, Alhamzah Alnoor, Noor S Baqer, Laith Alzubaidi, Osamah Shihab Albahri, Abdullah Hussein Alamoodi, Jinshuai Bai, Asma Salhi, et al. A systematic review of trustworthy and explainable artificial intelligence in healthcare: Assessment of quality, bias risk, and data fusion. *Information Fusion*, 96:156–191, 2023.

[12] Sajid Ali, Tamer Abuhmed, Shaker El-Sappagh, Khan Muhammad, Jose M Alonso-Moral, Roberto Confalonieri, Riccardo Guidotti, Javier Del Ser, Natalia Díaz-Rodríguez, and Francisco Herrera. Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence. *Information fusion*, 99:101805, 2023.

[13] Kasun Amarasinghe, Kit T Rodolfa, Sérgio Jesus, Valerie Chen, Vladimir Balayan, Pedro Saleiro, Pedro Bizarro, Ameet Talwalkar, and Rayid Ghani. On the importance of application-grounded experimental design for evaluating explainable ml methods. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 20921–20929, 2024.

[14] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in AI safety. *arXiv preprint arXiv:1606.06565*, 2016.

[15] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias. In *Ethics of data and analytics*, pages 254–264. Auerbach Publications, 2022.

[16] Adriana Anido-Alonso and Diego Alvarez-Estevez. Decentralized data-privacy preserving deep-learning approaches for enhancing inter-database generalization in automatic sleep staging. *IEEE Journal of Biomedical and Health Informatics*, 2023.

[17] Anna Markella Antoniadi, Yuhan Du, Yasmine Guendouz, Lan Wei, Claudia Mazo, Brett A Becker, and Catherine Mooney. Current challenges and future opportunities for XAI in machine learning-based clinical decision support systems: a systematic review. *Applied Sciences*, 11(11):5088, 2021.

[18] Jerome Argod, Jean-Louis Pepin, and Patrick Levy. Differentiating obstructive and central sleep respiratory events through pulse transit time. *American journal of respiratory and critical care medicine*, 158(6):1778–1783, 1998.

[19] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information fusion*, 58:82–115, 2020.

[20] Noor A Aziz, Awais Manzoor, Muhammad Deedahwar Mazhar Qureshi, M Atif Qureshi, and Wael Rashwan. Explainable AI in Healthcare: Systematic Review of Clinical Decision Support Systems. *medRxiv*, pages 2024–08, 2024.

[21] Niklas Babendererde, Amin Ranem, Moritz Fuchs, Camila Gonzalez, Henry John Krumb, and Anirban Mukhopadhyay. Fda's pccp: Opportunities and gaps. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 56–64. Springer, 2025.

[22] Hubert Baniecki and Przemyslaw Biecek. Adversarial attacks and defenses in explainable artificial intelligence: A survey. *Information Fusion*, 107:102303, 2024.

[23] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. Does the whole exceed its parts? the effect of ai explanations on complementary team performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–16, 2021.

[24] Sam Baron, Andrew J Latham, and Somogy Varga. Explainable ai and stakes in medicine: A user study. *Artificial Intelligence*, 340:104282, 2025.

[25] Rosemary Batt. Financialization through Health IT, Part I: Lessons from Electronic Health Systems. 2025.

[26] Richard B Berry, Rohit Budhiraja, Daniel J Gottlieb, David Gozal, Conrad Iber, Vishesh K Kapur, Carole L Marcus, Reena Mehra, Sairam Parthasarathy, Stuart F Quan, et al. Rules for scoring respiratory events in sleep: update of the 2007 AASM manual for the scoring of sleep and associated events: deliberations of the sleep apnea definitions task force of the American Academy of Sleep Medicine. *Journal of clinical sleep medicine*, 8(5):597–619, 2012.

[27] Bertrand. SweetViz. Visualize and compare datasets, target values and associations, with one line of code. https://github.com/fbdesignpro/sweetviz, 2020. Accessed: 2024-06-04.

[28] Clara Bove, Thibault Laugel, Marie-Jeanne Lesot, Charles Tijus, and Marcin Detyniecki. Why do explanations fail? A typology and discussion on failures in XAI. *arXiv preprint arXiv:2405.13474*, 2024.

[29] Sebastian Bruckert, Bettina Finzel, and Ute Schmid. The next generation of medical decision support: A roadmap toward transparent expert companions. *Frontiers in artificial intelligence*, 3:507973, 2020.

[30] Zana Buçinca, Phoebe Lin, Krzysztof Z Gajos, and Elena L Glassman. Proxy tasks and subjective measures can be misleading in evaluating explainable ai systems. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*, pages 454–464, 2020.

[31] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z Gajos. To trust or to think: cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. *Proceedings of the ACM on Human-computer Interaction*, 5(CSCW1):1–21, 2021.

[32] Chaofan Chen, Oscar Li, Daniel Tao, Alina Barnett, Cynthia Rudin, and Jonathan K Su. This looks like that: deep learning for interpretable image recognition. *Advances in neural information processing systems*, 32, 2019.

[33] Daphne Chylinski, Franziska Rudzik, Dorothée Coppieters ‘t Wallant, Martin Grignard, Nora Vandeleene, Maxime Van Egroo, Laurie Thiesse, Stig Solbach, Pierre Maquet, Christophe Phillips, et al. Validation of an automatic arousal detection algorithm for whole-night sleep EEG recordings. *Clocks & sleep*, 2(3):258–272, 2020.

[34] William J Clancey. The epistemology of a rule-based expert system—a framework for explanation. *Artificial intelligence*, 20(3):215–251, 1983.

[35] Eoin Delaney, Derek Greene, and Mark T Keane. Instance-based counterfactual explanations for time series classification. In *International conference on case-based reasoning*, pages 32–47. Springer, 2021.

[36] Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.

[37] Franz Ehrlich, Tony Sehr, Moritz Brandt, Martin Schmidt, Hagen Malberg, Martin Sedlmayr, and Miriam Goldammer. State-of-the-art sleep arousal detection evaluated on a comprehensive clinical dataset. *Scientific Reports*, 14(1):16239, 2024.

[38] Issam El Naqa and Martin J Murphy. What is machine learning? In *Machine learning in radiation oncology: theory and applications*, pages 3–11. Springer, 2015.

[39] Lorenzo Famiglini, Andrea Campagner, Marilia Barandas, Giovanni Andrea La Maida, Enrico Gallazzi, and Federico Cabitza. Evidence-based XAI: An empirical approach to design more effective and explainable decision support systems. *Computers in biology and medicine*, 170: 108042, 2024.

[40] Robert Fonod. DeepSleep 2.0: automated sleep arousal segmentation via deep learning. *AI*, 3 (1):164–179, 2022.

[41] U.S. Food and Drug Administration. Good Machine Learning Practice for Medical Device Development: Guiding Principles, October 2021. URL https://www.fda.gov/media/153486/download. Accessed: 2025-10-19.

[42] U.S. Food and Drug Administration. Transparency for Machine Learning-Enabled Medical Devices: Guiding Principles, June 2024. URL https://www.fda.gov/media/179269/download?attachment. Accessed: 2025-10-19.

[43] Andia Foroughi, Fardad Farokhi, Fereidoun Nowshiravan Rahatabad, and Alireza Kashaninia. Deep convolutional architecture-based hybrid learning for sleep arousal events detection through single-lead EEG signals. *Brain and behavior*, 13(6):e3028, 2023.

[44] Karl A Franklin and Eva Lindberg. Obstructive sleep apnea is a common disorder in the population—a review on the epidemiology of sleep apnea. *Journal of thoracic disease*, 7(8): 1311, 2015.

[45] Timo Freiesleben and Gunnar König. Dear XAI community, we need to talk! Fundamental misconceptions in current XAI research. In *World Conference on Explainable Artificial Intelligence*, pages 48–65. Springer, 2023.

[46] Alex A Freitas. Comprehensible classification models: a position paper. *ACM SIGKDD explorations newsletter*, 15(1):1–10, 2014.

[47] Benjamin Fresz, Elena Dubovitskaya, Danilo Brajovic, Marco F Huber, and Christian Horz. How should ai decisions be explained? requirements for explanations from the perspective of european law. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, pages 438–450, 2024.

[48] Azul Garza and Max Mergenthaler-Canseco. TimeGPT-1. *arXiv preprint arXiv:2310.03589*, 2023.

[49] Susanne Gaube, Harini Suresh, Martina Raue, Alexander Merritt, Seth J Berkowitz, Eva Lermer, Joseph F Coughlin, John V Guttag, Errol Colak, and Marzyeh Ghassemi. Do as ai say: susceptibility in deployment of clinical decision-aids. *NPJ digital medicine*, 4(1):31, 2021.

[50] Susanne Gaube, Harini Suresh, Martina Raue, Eva Lermer, Timo K Koch, Matthias FC Hudecek, Alun D Ackery, Samir C Grover, Joseph F Coughlin, Dieter Frey, et al. Non-task expert physicians benefit from correct explainable AI advice when reviewing X-rays. *Scientific reports*, 13(1):1383, 2023.

[51] Jin Ge, Valy Fontil, Sara Ackerman, Mark J Pletcher, and Jennifer C Lai. Clinical decision support and electronic interventions to improve care quality in chronic liver diseases and cirrhosis. *Hepatology*, 81(4):1353–1364, 2025.

[52] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92, 2021.

[53] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

[54] Marc D Gellman. *Encyclopedia of behavioral medicine*. Springer, 2020.

[55] Marzyeh Ghassemi, Leo Anthony Celi, and David J Stone. State of the art review: the data revolution in critical care. *Annual Update in Intensive Care and Emergency Medicine 2015*, pages 573–586, 2015.

[56] Mohammad M Ghassemi, Benjamin E Moody, Li-Wei H Lehman, Christopher Song, Qiao Li, Haoqi Sun, Roger G Mark, M Brandon Westover, and Gari D Clifford. You snooze, you win: the physionet/computing in cardiology challenge 2018. In *2018 Computing in Cardiology Conference (CinC)*, volume 45, pages 1–4. IEEE, 2018.

[57] Leilani H Gilpin, David Bau, Ben Z Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. Explaining explanations: An overview of interpretability of machine learning. In *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*, pages 80–89. IEEE, 2018.

[58] Ary L Goldberger, Luis AN Amaral, Leon Glass, Jeffrey M Hausdorff, Plamen Ch Ivanov, Roger G Mark, Joseph E Mietus, George B Moody, Chung-Kang Peng, and H Eugene Stanley. PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. *circulation*, 101(23):e215–e220, 2000.

[59] Bryce Goodman and Seth Flaxman. European Union regulations on algorithmic decision-making and a "right to explanation". *AI magazine*, 38(3):50–57, 2017.

[60] WHO Guidance. Ethics and governance of artificial intelligence for health. *World Health Organization*, 2021.

[61] David Gunning. Explainable artificial intelligence (xai). *Defense advanced research projects agency (DARPA), nd Web*, 2(2):1, 2017.

[62] David Gunning and David Aha. DARPA's explainable artificial intelligence (XAI) program. *AI magazine*, 40(2):44–58, 2019.

[63] David Gunning, Eric Vorm, Yunyan Wang, and Matt Turek. DARPA's explainable AI (XAI) program: A retrospective. *Authorea Preprints*, 2021.

[64] Jyoti Gupta and KR Seeja. A comparative study and systematic analysis of XAI models and their applications in healthcare. *Archives of Computational Methods in Engineering*, 31(7): 3977–4002, 2024.

[65] P Haettenschwiler. Neues anwenderfreundliches Konzept der Entscheidungs-unterstutzung. Gutes Entscheiden in Wirtschaft, Politik und Gesellschaft. Zurich: Hochschulverlag AG, 1999. *S. 189*, 208, 1999.

[66] Ronan Hamon, Henrik Junklewitz, Ignacio Sanchez, Gianclaudio Malgieri, and Paul De Hert. Bridging the gap between ai and explainability in the gdpr: towards trustworthiness-by-design in automated decision-making. *IEEE Computational Intelligence Magazine*, 17(1):72–85, 2022.

[67] Andreas Holzinger, Chris Biemann, Constantinos S Pattichis, and Douglas B Kell. What do we need to build explainable AI systems for the medical domain? *arXiv preprint arXiv:1712.09923*, 2017.

[68] Andreas Holzinger, André Carrington, and Heimo Müller. Measuring the quality of explanations: the system causability scale (SCS) comparing human and machine explanations. *KI-Künstliche Intelligenz*, 34(2):193–198, 2020.

[69] Jeonghwan Hwang, Taeheon Lee, Honggu Lee, and Seonjeong Byun. A clinical decision support system for sleep staging tasks with explanations from artificial intelligence: user-centered design and evaluation study. *Journal of medical Internet research*, 24(1):e28659, 2022.

[70] Syed Anas Imtiaz. A systematic review of sensing technologies for wearable sleep staging. *Sensors*, 21(5):1562, 2021.

[71] Jinsun Jung, Sunghoon Kang, Jeeyae Choi, Robert El-Kareh, Hyungbok Lee, and Hyeoneui Kim. Evaluating the impact of explainable ai on clinicians' decision-making: A study on ICU length of stay prediction. *International Journal of Medical Informatics*, page 105943, 2025.

[72] Mark T Keane and Eoin M Kenny. How case-based reasoning explains neural networks: A theoretical analysis of XAI using post-hoc explanation-by-example from a survey of ANN-CBR twin-systems. In *International conference on case-based reasoning*, pages 155–171. Springer, 2019.

[73] Georgios Kostopoulos, Gregory Davrazos, and Sotiris Kotsiantis. Explainable Artificial Intelligence-Based Decision Support Systems: A Recent Review. *Electronics*, 13(14):2842, 2024.

[74] Stefan Kraft, Klaus Broelemann, Andreas Theissler, and Gjergji Kasneci. SPARROW: Semantically Coherent Prototypes for Image Classification. In *BMVC*, page 186, 2021.

[75] Stefan Kraft, Andreas Theissler, Vera Wienhausen-Wilke, Philipp Walter, and Gjergji Kasneci. Comprehensive Polysomnography (CPS) Dataset: A Resource for Sleep-Related Arousal

Research. PhysioNet data repository, 2024. URL https://doi.org/10.13026/sxs0-h317. Dataset.

[76] Stefan Kraft, Andreas Theissler, Vera Wienhausen-Wilke, Gjergji Kasneci, and Hendrik Lensch. Assessing the real-world utility of explainable ai for arousal diagnostics: An application-grounded user study. *arXiv preprint arXiv:2510.21389*, 2025.

[77] Stefan Kraft, Andreas Theissler, Vera Wienhausen-Wilke, Philipp Walter, Gjergji Kasneci, and Hendrik Lensch. ALPEC: A comprehensive evaluation framework and dataset for machine learning-based arousal detection in clinical practice. In Xuhai Orson Xu, Edward Choi, Pankhuri Singhal, Walter Gerych, Shengpu Tang, Monica Agrawal, Adarsh Subbaswamy, Elena Sizikova, Jessilyn Dunn, Roxana Daneshjou, Tasmie Sarker, Matthew McDermott, and Irene Chen, editors, *Conference on Health, Inference, and Learning, UC Berkeley, Berkeley, USA, 25-27 June 2025*, volume 287 of *Proceedings of Machine Learning Research*, pages 395–429. PMLR, 2025. URL https://proceedings.mlr.press/v287/kraft25a.html.

[78] Todd Kulesza, Simone Stumpf, Margaret Burnett, Sherry Yang, Irwin Kwan, and Weng-Keen Wong. Too much, too little, or just right? Ways explanations impact end users' mental models. In *2013 IEEE Symposium on visual languages and human centric computing*, pages 3–10. IEEE, 2013.

[79] Vivian Lai, Chacha Chen, Q Vera Liao, Alison Smith-Renner, and Chenhao Tan. Towards a science of human-ai decision making: a survey of empirical studies. *arXiv preprint arXiv:2112.11471*, 2021.

[80] Tobias Leemann, Yao Rong, Stefan Kraft, Enkelejda Kasneci, and Gjergji Kasneci. Coherence evaluation of visual concepts with objects and language. In *ICLR2022 Workshop on the Elements of Reasoning: Objects, Structure and Causality*, 2022.

[81] Hongyang Li and Yuanfang Guan. DeepSleep convolutional neural network allows accurate and fast detection of sleep arousal. *Communications biology*, 4(1):18, 2021.

[82] Oscar Li, Hao Liu, Chaofan Chen, and Cynthia Rudin. Deep learning for case-based reasoning through prototypes: A neural network that explains its predictions. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

[83] Rensis Likert. A technique for the measurement of attitudes. *Archives of psychology*, 1932.

[84] Pantelis Linardatos, Vasilis Papastefanopoulos, and Sotiris Kotsiantis. Explainable ai: A review of machine learning interpretability methods. *Entropy*, 23(1):18, 2020.

[85] Zachary C Lipton. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 16(3):31–57, 2018.

[86] Luca Longo, Mario Brcic, Federico Cabitza, Jaesik Choi, Roberto Confalonieri, Javier Del Ser, Riccardo Guidotti, Yoichi Hayashi, Francisco Herrera, Andreas Holzinger, et al. Explainable Artificial Intelligence (XAI) 2.0: A manifesto of open challenges and interdisciplinary research directions. *Information Fusion*, 106:102301, 2024.

[87] Stella Lowry and Gordon Macpherson. A blot on the profession. *British medical journal (Clinical research ed.)*, 296(6623):657, 1988.

[88] Scott Lundberg. A unified approach to interpreting model predictions. *arXiv preprint arXiv:1705.07874*, 2017.

[89] Yasmine Madan, Argyrios Perivolaris, Robert Chris Adams-McGavin, and James J Jung. Clinician interaction with artificial intelligence systems: a narrative review. *Journal of Medical Artificial Intelligence*, 2024.

[90] Tambiama Madiega. Artificial intelligence act. *European Parliament: European Parliamentary Research Service*, 2021.

[91] Usman Mahmood, Robik Shrestha, David DB Bates, Lorenzo Mannelli, Giuseppe Corrias, Yusuf Emre Erdi, and Christopher Kanan. Detecting spurious correlations with sanity tests for artificial intelligence guided radiology systems. *Frontiers in digital health*, 3:671015, 2021.

[92] Joshua C Mandel, David A Kreda, Kenneth D Mandl, Isaac S Kohane, and Rachel B Ramoni. SMART on FHIR: a standards-based, interoperable apps platform for electronic health records. *Journal of the American Medical Informatics Association*, 23(5):899–908, 2016.

[93] Cindy Marling, Stefania Montani, Isabelle Bichindaritz, and Peter Funk. Synergistic case-based reasoning in medical domains. *Expert systems with applications*, 41(2):249–259, 2014.

[94] Joao Marques-Silva and Xuanxiang Huang. Explainability is NOT a game. *Communications of the ACM*, 67(7):66–75, 2024.

[95] John McCarthy. Programs with common sense, 1959.

[96] Stephanie L McNamara, Paul H Yi, and William Lotter. The clinician-AI interface: intended use and explainability in FDA-cleared AI devices for medical image interpretation. *NPJ Digital Medicine*, 7(1):80, 2024.

[97] Melkamu Mersha, Khang Lam, Joseph Wood, Ali AlShami, and Jugal Kalita. Explainable artificial intelligence: A survey of needs, techniques, applications, and future direction. *Neurocomputing*, page 128111, 2024.

[98] Blackford Middleton, DF Sittig, and A Wright. Clinical decision support: a 25 year retrospective and a 25 year vision. *Yearbook of medical informatics*, 25(S 01):S103–S116, 2016.

[99] George A Miller. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological review*, 63(2):81, 1956.

[100] Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*, 267:1–38, 2019.

[101] Tim Miller, Piers Howe, and Liz Sonenberg. Explainable AI: Beware of inmates running the asylum or: How I learnt to stop worrying and love the social and behavioural sciences. *arXiv preprint arXiv:1712.00547*, 2017.

[102] Yao Ming, Panpan Xu, Huamin Qu, and Liu Ren. Interpretable and steerable sequence learning via prototypes. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 903–913, 2019.

[103] Riccardo Miotto, Li Li, Brian A Kidd, and Joel T Dudley. Deep patient: an unsupervised representation to predict the future of patients from the electronic health records. *Scientific reports*, 6(1):1–10, 2016.

[104] Tomofumi Misaka, Yuko Niimura, Akiomi Yoshihisa, Kento Wada, Yusuke Kimishima, Tetsuro Yokokawa, Satoshi Abe, Masayoshi Oikawa, Takashi Kaneshiro, Atsushi Kobayashi, et al. Clinical impact of sleep-disordered breathing on very short-term blood pressure variability determined by pulse transit time. *Journal of Hypertension*, 38(9):1703–1711, 2020.

[105] Sina Mohseni, Niloofar Zarei, and Eric D Ragan. A multidisciplinary survey and framework for design and evaluation of explainable AI systems. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 11(3-4):1–45, 2021.

[106] Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. Methods for interpreting and understanding deep neural networks. *Digital signal processing*, 73:1–15, 2018.

[107] Shane T Mueller, Robert R Hoffman, William Clancey, Abigail Emrey, and Gary Klein. Explanation in human-AI systems: A literature meta-review, synopsis of key ideas and publications, and bibliography for explainable AI. *arXiv preprint arXiv:1902.01876*, 2019.

[108] Meike Nauta, Annemarie Jutte, Jesper Provoost, and Christin Seifert. This looks like that, because... explaining prototypes for interpretable image recognition. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 441–456. Springer, 2021.

[109] Meike Nauta, Jan Trienes, Shreyasi Pathak, Elisa Nguyen, Michelle Peters, Yasmin Schmitt, Jörg Schlötterer, Maurice Van Keulen, and Christin Seifert. From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable ai. *ACM Computing Surveys*, 55(13s):1–42, 2023.

[110] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453, 2019.

[111] Damilola Osamika, Bamidele Samuel Adelusi, Maria Theresa Chinyeaka Kelvin-Agwu, Ashiata Yetunde Mustapha, Adelaide Yeboah Forkuo, and Nura Ikhalea. A critical review of health data interoperability standards: Fhir, hl7, and beyond.

[112] Jerome A Osheroff, Jonathan M Teich, Blackford Middleton, Elaine B Steen, Adam Wright, and Don E Detmer. A roadmap for national action on clinical decision support. *Journal of the American medical informatics association*, 14(2):141–145, 2007.

[113] Cecilia Panigutti, Andrea Beretta, Fosca Giannotti, and Dino Pedreschi. Understanding the impact of explanations on advice-taking: a user study for AI-based clinical Decision Support Systems. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pages 1–9, 2022.

[114] Cecilia Panigutti, Ronan Hamon, Isabelle Hupont, David Fernandez Llorca, Delia Fano Yela, Henrik Junklewitz, Salvatore Scalzo, Gabriele Mazzini, Ignacio Sanchez, Josep Soler Garrido, et al. The role of explainable AI in the context of the AI Act. In *Proceedings of the 2023 ACM conference on fairness, accountability, and transparency*, pages 1139–1150, 2023.

[115] Petros Papadopoulos, Mario Soflano, Yaelle Chaudy, Wilson Adejo, and Thomas M Connolly. A systematic review of technologies and standards used in the development of rule-based clinical decision support systems. *Health and Technology*, 12(4):713–727, 2022.

[116] Shreyasi Pathak, Jörg Schlötterer, Jeroen Veltman, Jeroen Geerdink, Maurice van Keulen, and Christin Seifert. Prototype-based interpretable breast cancer prediction models: Analysis and challenges. In *World Conference on Explainable Artificial Intelligence*, pages 21–42. Springer, 2024.

[117] Henna Pitkänen, Sami Nikkonen, Marika Rissanen, Anna Sigridur Islind, Heidur Gretarsdottir, Erna Sif Arnardottir, Timo Leppänen, and Henri Korkalainen. Multi-centre arousal scoring agreement in the Sleep Revolution. *Journal of Sleep Research*, 33(4):e14127, 2024.

[118] DJ Pitson et al. Value of beat-to-beat blood pressure changes, detected by pulse transit time, in the management of the obstructive sleep apnoea/hypopnoea syndrome. *European Respiratory Journal*, 12(3):685–692, 1998.

[119] Forough Poursabzi-Sangdeh, Daniel G Goldstein, Jake M Hofman, Jennifer Wortman Wortman Vaughan, and Hanna Wallach. Manipulating and measuring model interpretability. In *Proceedings of the 2021 CHI conference on human factors in computing systems*, pages 1–52, 2021.

[120] Daniel J Power. *Decision support systems: concepts and resources for managers*, volume 13. Quorum Books Westport, 2002.

[121] Naresh M Punjabi. The epidemiology of adult obstructive sleep apnea. *Proceedings of the American Thoracic Society*, 5(2):136–143, 2008.

[122] F Raschke and J Fischer. "Arousal" in der Schlafmedizin. *Somnologie*, 1(2), 1997.

[123] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Why should i trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD)*. ACM, 2016.

[124] Emmanouil S Rigas, Paris Lagakis, Makis Karadimas, Evangelos Logaras, Dimitra Latsou, Magda Hatzikou, Athanasios Poulakidas, Antonis Billis, and Panagiotis D Bamidis. Semantic interoperability for an AI-based applications platform for smart hospitals using HL7 FHIR. *Journal of Systems and Software*, 215:112093, 2024.

[125] Thomas Rojat, Raphaël Puget, David Filliat, Javier Del Ser, Rodolphe Gelin, and Natalia Díaz-Rodríguez. Explainable artificial intelligence (xai) on timeseries data: A survey. *arXiv preprint arXiv:2104.00950*, 2021.

[126] Yao Rong, Tobias Leemann, Thai-trang Nguyen, Lisa Fiedler, Peizhu Qian, Vaibhav Unhelkar, Tina Seidel, Gjergji Kasneci, and Enkelejda Kasneci. Towards human-centered explainable AI: user studies for model explanations. 2022.

[127] Rikard Rosenbacke, Åsa Melhus, Martin McKee, and David Stuckler. How explainable artificial intelligence can increase or decrease clinicians' trust in ai applications in health care: systematic review. *Jmir Ai*, 3:e53207, 2024.

[128] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence*, 1(5):206–215, 2019.

[129] Cynthia Rudin, Chaofan Chen, Zhi Chen, Haiyang Huang, Lesia Semenova, and Chudi Zhong. Interpretable machine learning: Fundamental principles and 10 grand challenges. *Statistic Surveys*, 16:1–85, 2022.

[130] Aurelia Sauerbrei, Angeliki Kerasidou, Federica Lucivero, and Nina Hallowell. The impact of artificial intelligence on the person-centred, doctor-patient relationship: some problems and solutions. *BMC Medical Informatics and Decision Making*, 23(1):73, 2023.

[131] Philipp Schmidt and Felix Biessmann. Calibrating human-ai collaboration: Impact of risk, ambiguity and transparency on algorithmic bias. In *Machine Learning and Knowledge Extraction: 4th IFIP TC 5, TC 12, WG 8.4, WG 8.9, WG 12.9 International Cross-Domain Conference, CD-MAKE 2020, Dublin, Ireland, August 25–28, 2020, Proceedings 4*, pages 431–449. Springer, 2020.

[132] Gesina Schwalbe and Bettina Finzel. A comprehensive taxonomy for explainable artificial intelligence: a systematic survey of surveys on methods and concepts. *Data Mining and Knowledge Discovery*, 38(5):3043–3101, 2024.

[133] Ruey-Kai Sheu and Mayuresh Sunil Pardeshi. A survey on medical explainable AI (XAI): recent progress, explainability approach, human interaction and scoring system. *Sensors*, 22 (20):8068, 2022.

[134] Edward H Shortliffe and Bruce G Buchanan. A model of inexact reasoning in medicine. *Mathematical biosciences*, 23(3-4):351–379, 1975.

[135] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *International conference on machine learning*, pages 3145–3153. PMlR, 2017.

[136] Ilija Šimić, Vedran Sabol, and Eduardo Veas. Xai methods for neural time series classification: A brief review. *arXiv preprint arXiv:2108.08009*, 2021.

[137] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

[138] Venkatesh Sivaraman, Leigh A Bukowski, Joel Levin, Jeremy M Kahn, and Adam Perer. Ignore, trust, or negotiate: understanding clinician acceptance of AI-based treatment recommendations in health care. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–18, 2023.

[139] Dylan Slack, Sophie Hilgard, Emily Jia, Sameer Singh, and Himabindu Lakkaraju. Fooling lime and shap: Adversarial attacks on post hoc explanation methods. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 180–186, 2020.

[140] Sondre Sørbø and Massimiliano Ruocco. Navigating the metric maze: A taxonomy of evaluation metrics for anomaly detection in time series. *Data Mining and Knowledge Discovery*, pages 1–42, 2023.

[141] Francesco Sovrano and Fabio Vitali. Perlocution vs illocution: How different interpretations of the act of explaining impact on the evaluation of explanations and xai. In *World Conference on Explainable Artificial Intelligence*, pages 25–47. Springer, 2023.

[142] Francesco Sovrano, Fabio Vitali, and Monica Palmirani. Modelling gdpr-compliant explanations for trustworthy ai. In *International Conference on Electronic Government and the Information Systems Perspective*, pages 219–233. Springer, 2020.

[143] Reed T Sutton, David Pincock, Daniel C Baumgart, Daniel C Sadowski, Richard N Fedorak, and Karen I Kroeker. An overview of clinical decision support systems: benefits, risks, and strategies for success. *NPJ digital medicine*, 3(1):17, 2020.

[144] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

[145] Andreas Theissler, Francesco Spinnato, Udo Schlegel, and Riccardo Guidotti. Explainable AI for time series classification: a review, taxonomy and research directions. *Ieee Access*, 10: 100700–100724, 2022.

[146] Sana Tonekaboni, Shalmali Joshi, Melissa D McCradden, and Anna Goldenberg. What clinicians want: contextualizing explainable machine learning for clinical end use. In *Machine learning for healthcare conference*, pages 359–380. PMLR, 2019.

[147] Eric J Topol. High-performance medicine: the convergence of human and artificial intelligence. *Nature medicine*, 25(1):44–56, 2019.

[148] U.S. Food and Drug Administration. Artificial Intelligence-Enabled Medical Devices, 2025. URL https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-enabled-medical-devices. Accessed: 2025-10-22.

[149] Sven Van Laere, Katoo M Muylle, and Pieter Cornu. Clinical decision support and new regulatory frameworks for medical devices: are we ready for it?-a viewpoint paper. *International Journal of Health Policy and Management*, 11(12):3159, 2021.

[150] Michael Van Lent, William Fisher, and Michael Mancuso. An explainable artificial intelligence system for small-unit tactical behavior. In *Proceedings of the national conference on artificial intelligence*, pages 900–907. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2004.

[151] Simon Vollert, Martin Atzmueller, and Andreas Theissler. Interpretable Machine Learning: A brief survey from the predictive maintenance perspective. In *2021 26th IEEE international conference on emerging technologies and factory automation (ETFA)*, pages 01–08. IEEE, 2021.

[152] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011.

[153] Dakuo Wang, Liuping Wang, Zhan Zhang, Ding Wang, Haiyi Zhu, Yvonne Gao, Xiangmin Fan, and Feng Tian. “Brilliant AI doctor” in rural clinics: Challenges in AI-powered clinical

decision support system deployment. In *Proceedings of the 2021 CHI conference on human factors in computing systems*, pages 1–18, 2021.

[154] Xinru Wang and Ming Yin. Are explanations helpful? a comparative study of the effects of explanations in ai-assisted decision-making. In *26th International Conference on Intelligent User Interfaces*, pages 318–328, 2021.

[155] Thomas-Christian Wetter, Roland Popp, Michael Arzt, and Thomas Pollmächer. *ELSEVIER ESSENTIALS Schlafmedizin: Das Wichtigste für Ärzte aller Fachrichtungen*. Elsevier Health Sciences, 2012.

[156] Vera Wienhausen-Wilke and Stefan Kraft. Computer-aided diagnostics of sleep-related arousals on the basis of pulse wave analyses, 2024. URL https://drks.de/search/en/trial/DRKS00033641. Accessed: 2025-03-31.

[157] Tom Nuno Wolf, Fabian Bongratz, Anne-Marie Rickmann, Sebastian Pölsterl, and Christian Wachinger. Keep the faith: Faithful explanations in convolutional neural networks for case-based reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 5921–5929, 2024.

[158] Qian Xu, Wenzhao Xie, Bolin Liao, Chao Hu, Lu Qin, Zhengzijin Yang, Huan Xiong, Yi Lyu, Yue Zhou, and Aijing Luo. Interpretability of clinical decision support systems based on artificial intelligence from technological and medical perspective: A systematic review. *Journal of healthcare engineering*, 2023(1):9919269, 2023.

[159] Yunguo Yu, Cesar A Gomez-Cabello, Syed Ali Haider, Ariana Genovese, Srinivasagam Prabha, Maissa Trabilsy, Bernardo G Collaco, Nadia G Wood, Sanjay Bagaria, Cui Tao, et al. Enhancing clinician trust in ai diagnostics: A dynamic framework for confidence calibration and transparency. *Diagnostics*, 15(17):2204, 2025.

[160] Jingwei Zhang, Xiaodong Yang, Yiqiang Chen, and Ruizhe Sun. C-PPT: A Channel-Wise Prototypical Part Transformer for Interpretable Perioperative Complication Prediction with Blood Pressure. In *International Conference on Pattern Recognition*, pages 46–60. Springer, 2024.

[161] Zhijie Zhu, Lei Fan, Maurice Pagnucco, and Yang Song. Interpretable Image Classification via Non-parametric Part Prototype Learning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 9762–9771, 2025.

A SPARROW: Semantically Coherent Prototypes for Image Classification

The full text of the following publication is included in this appendix:

Stefan Kraft, Klaus Broelemann, Andreas Theissler, and Gjergji Kasneci. SPARROW: Semantically Coherent Prototypes for Image Classification. In *BMVC*, page 186, 2021

SPARROW: Semantically Coherent Prototypes for Image Classification

Stefan Kraft[1,4]
stefan.kraft@stz-softwaretechnik.de
Klaus Broelemann[2]
klaus.broelemann@schufa.de
Andreas Theissler[3]
https://orcid.org/0000-0003-0746-0424
Gjergji Kasneci[4,2]
gjergji.kasneci@uni-tuebingen.de

[1] IT-Designers Group
Esslingen am Neckar, GER
[2] SCHUFA Holding AG
Wiesbaden, GER
[3] Aalen University of Applied Sciences
Aalen, GER
[4] Data Science & Analytics Research
The University of Tübingen
Tübingen, GER

Abstract

Current prototype-based classification often leads to prototypes with overlapping semantics where several prototypes are similar to the same image parts. Also, single prototypes tend to activate highly on a mixture of semantically different image parts. This impedes interpretability since the nature of the connections between the parts is unknown. We propose a framework that is comprised of two key elements: (i) A novel method which leads to semantically coherent prototypes and (ii) an evaluation protocol which is based on part annotations and allows to quantitatively compare the explanatory capacity of prototypes from different methods. We demonstrate the viability of our framework by comparing our method to a standard prototype-based classification method and show that our method is capable of producing prototypes of superior interpretability.

1 Introduction

Recent research has called for ML models that are interpretable by design [20] rather than post-hoc explanations on black-box models. However, in image processing most explanation approaches rely on activation and saliency maps, which can be highly misleading and sometimes even lead to spurious saliency maps [1].

Prototype classification, that is classifying samples based on their similarity to prototypical samples in a latent space, aims to achieve this kind of interpretability [20]. However, there is a performance-interpretability trade-off with regard to the number of prototypes [6]. While the classification performance typically increases with an increasing number of prototypes, the explanation quality for a local prediction diminishes; even more so if either prototypes do not relate to specific semantic concepts or if they overlap on the same semantic concepts.

This work complements the literature on prototype classification by suggesting a novel framework for prototype quality: SPARROW (uniqueneSs – sPArsity – naRROWness). SPARROW

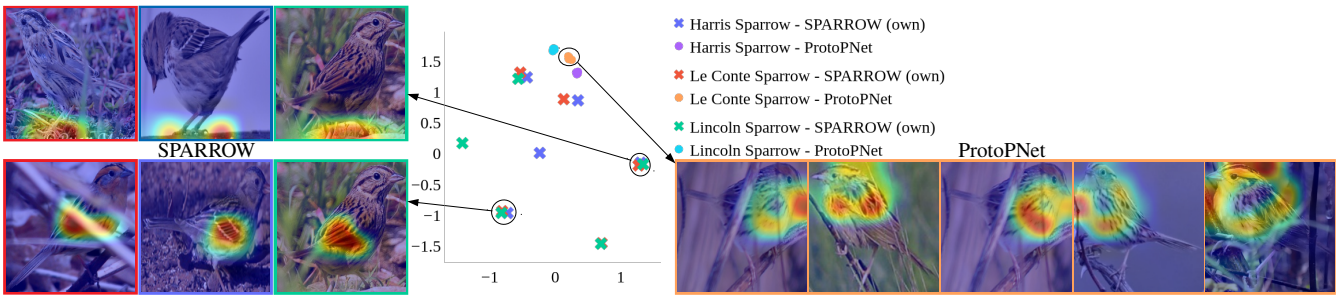


Figure 1: **Comparison of prototypes** for ProtoPNet [6] and our own method: SPARROW. The latter encourages semantically coherent prototypes (*e.g.* legs, wing) while ProtoPNet selects class prototypes (*e.g.* Le Conte Sparrow). The center panel shows a combined t-SNE visualization of the latent prototype spaces based on a cosine distance measure. Samples that represent the prototypes are displayed (left: SPARROW, right: ProtoPNet) and connected to their coordinates in t-SNE space. The superior interpretability of prototypes from SPARROW will be quantified globally by means of our novel evaluation protocol.

enables researchers and practitioners alike to learn semantically coherent prototypes and evaluate their explanatory capacity. It is inspired by well understood interpretability principles: (1) **Sparsity** – Sparse explanations allow humans to understand how a few different concepts jointly form a model prediction [20]. (2) **Narrowness** – An explanation should capture a concept as narrowly as possible. *E.g.*, saliency areas covering larger image parts can be ambiguous and the observer can only guess what the pivotal factor in favor of the model's decision was [20]. (3) **Uniqueness** – The overlapping of concepts should be minimized.

We showcase the effectiveness of SPARROW – in terms of a quantifiable performance (instead of just visual inspection) and in terms of semantically coherent prototype generation – in comparison to state-of-the-art prototype generation [6].

ProtoPNet We briefly introduce ProtoPNet [6] on which the SPARROW prototype learning method is based. ProtoPNet utilizes a set of loss components (cf. table 1) to jointly learn neural network weights and weights of prototype vectors. It calculates similarities between parts (patches) of latent space sample embeddings and prototypes. Similarity scores are then used in a weighted superposition in a final fully-connected layer to derive class prediction scores. Since the similarities are not derived post-hoc and the last layer is simple and transparent, ProtoPNet qualifies as being intrinsically interpretable. During training, fixed and evenly distributed class identities for prototypes are enforced. In the end, prototypes are projected onto the latent space patch of a training sample which they are most similar to. Thus, prototypes can naturally represent a part of a sample image and the latent space self-activation maps of prototypes can be upsampled and visualized in the input space. The same can be done for activation maps between prototypes and test samples so that the reason for their similarity can be visually inspected. Details about the model and training process are available in the supplementary material which also contains a table of notation.

Case Study We want to demonstrate that semantically coherent prototypes lead to less ambiguous visual interpretability. Figure 1 shows upsampled latent space self-activation maps of prototypes from ProtoPNet [6] and our method (SPARROW). While it is difficult for the competitor to put semantic labels on the self-activation maps of prototypes, because

Component	Goal	Type	Weight
$\mathcal{L}^{\text{CrsEnt}}$	Classification performance	CE	1
$\mathcal{L}^{\text{Clst}}$	Cluster samples around prototypes	Agg	1
$\mathcal{L}^{\text{Sep}}$	Separate clusters by class	Agg	0
$\mathcal{L}^{\text{AS}*}$	Decorrelate prototypes	log-loss	1
$\mathcal{L}^{\text{PSD}*}$	Keep prototypes close to a sample	log-loss	100

Table 1: **Overview of loss components**. Components marked with * are our own addition to ProtoPNet [6]. CE: Cross Entropy, Agg: Aggregate function. Our choice of weights is discussed in section 3.

they are neither narrow nor unique, the prototypes from SPARROW are easier to interpret and less ambiguous. *E.g.* the activations of the bottom three images on the left seem to focus on the right wing while we could not make a similarly clear statement for the prototypes from ProtoPNet. Additionally, the t-SNE visualization in the middle shows different samples in cosine distance space. We consider cosine similarity as it serves as a natural measure to quantify correlations between prototypes. Prototypes from ProtoPNet are very similar to one another since they are closely clustered while our prototypes are further apart, indicating that they are pointing to different concepts. As Adebayo *et al.* demonstrated [1], simply judging two methods from a handful of samples is not sufficient to compare them with respect to their general capacity to produce semantically meaningful explanations, which makes the need for our SPARROW quantitative evaluation protocol all the more urgent.

Related Work *Interpretable Models by Design*: In a seminal position paper, Rudin [20] argued in favour of building predictive models that are explainable by design. Recent work followed this suggestion focusing on rendering neural networks interpretable using piecewise linear functions [2] or prototypes [6, 10, 25]. Our present work is most similar to the work of [6]. However, it differs from [6] by introducing novel loss components that encourage unique and narrow semantically coherent prototypes.

Interpretability measures: To the best of our knowledge other works in the field of prototype classification have not evaluated the interpretability of latent space prototype activations quantitatively but instead rely on qualitative visual assessments. In the broader field of predictions with convolutional neural networks (CNNs) Bau *et al.* [4] quantify the alignment of CNN units with various pixel-wise labeled concepts by performing binary segmentation on activation maps and calculating the intersection over union with labeled concepts. We also match activation maps with labels but use samples with keypoint part annotations instead of pixel-wise annoation maps. They proceed to count the number of distinct visual concepts that are matched per CNN layer. We follow a similar goal by measuring the completeness of part annotations that are matched by prototypes over all samples. Zhang *et al.* [24] extended this work by adding a location instability metric which measures the degree to which the inferred position of a CNN activation pattern in the input space varies with respect to a set of landmark positions over different images. We calculate a prototype focus measure which does not use other landmark positions but instead determines how consistent prototypes are in recurrently matching the same parts over all samples.

2 SPARROW

Method for Learning Semantically Coherent Prototypes In order to learn semantically coherent prototypes, we extend ProtoPNet [6] with two additional loss components. First, we add an angular similarity (*AS*) based loss which aids in decorrelating the prototypes:

$$\mathcal{L}^{\text{AS}} = -\frac{1}{C}\sum_{v=1}^{C}\max_{i,j\in\mathcal{I}_v}\log(1-\text{AS}(\mathbf{p}_i,\mathbf{p}_j)), \tag{1}$$

where $\mathcal{I}_v$ denotes the set of prototype indices of intra-class prototypes and C is the number of classes. The *AS* between prototypes $\mathbf{p}_i$ and $\mathbf{p}_j$ is [22] $\text{AS}(\mathbf{p}_i,\mathbf{p}_j) = 1 - \frac{1}{\pi}\arccos(\text{CS}(\mathbf{p}_i,\mathbf{p}_j))$ with $i,j \in \mathcal{I}_v$ and the cosine similarity (*CS*) is [19] $\text{CS}(\mathbf{p}_i,\mathbf{p}_j) = \frac{\mathbf{p}_i^T\cdot\mathbf{p}_j}{\|\mathbf{p}_i\|\|\mathbf{p}_j\|}$. The *AS* can be interpreted as a probability which allows to calculate a log-loss. As our initial experiments have shown, taking the maximum in eq. 1 reduces the variance of the angular similarity between intra-class prototype combinations as compared to other aggregate functions like average or sum. We also found that the optimization typically results in one prototype per class being close to samples of the class while the other prototypes become outliers in latent space. In order to fulfill the requirement by Chen *et al.* [6] to keep prototypes sufficiently close to samples in latent space, we implement a second loss component:

$$\mathcal{L}^{\text{PSD}} = -\frac{1}{m}\sum_{j=1}^{m}\log(1-\frac{\text{PSD}_j(X,\mathbf{p}_j)}{\text{dist}_{\max}}). \tag{2}$$

Here, m denotes the number of prototypes, $\text{dist}_{\max}$ is the maximum possible distance in latent space and the term that we call prototype-sample distance (*PSD*) is taken from Li *et al.* [10] as $\text{PSD}_j(X,\mathbf{p}_j) = \min_{\mathbf{z}\in\text{patches}(f(\mathbf{x}_\mu))}\|\mathbf{p}_j - \mathbf{z}\|^2$. It is calculated over all samples of the current batch, i.e. $\mathbf{x}_\mu \in X$ and $\mu \in \text{batch}([1,\ldots,n])$ with the set of training samples X of length n. Both new loss components are added to the total loss with static weights (cf. table 1).

Evaluation Protocol We propose an evaluation protocol which leverages information from part annotations to provide *explanatory capacity* estimates. On a high level, our approach is based on matching part annotations with activation masks. For each sample we assume that there are T keypoint annotations of sample parts available. The activation masks stem from the latent space activations of samples by prototypes which have the same class identity as the samples. These activation maps are upsampled to the input space and cropped to masks by an activation threshold. Choices for this threshold will be discussed in section 3. We show schematically annotated parts and prototype activation masks in figure 2. In the following we will refer to matches between activation masks and part annotations as “matches between prototypes and parts". If a prototype does not match any part, we select the closest part coordinate to the activation mask as a match. This is done because some activation masks are too narrow to match any part. For a more detailed explanation of the matching procedure we refer to the supplementary material. We are now ready to derive evaluation measures.

The decorrelation measure quantifies to which degree prototypes with the same class identity activate highly only on non-overlapping annotated part semantics. *E.g.* two prototypes that both activate highly on the wing of a bird have a low decorrelation score. Highly decorrelated prototypes typically lead to narrower saliency maps and therefore enable less ambiguous interpretations.

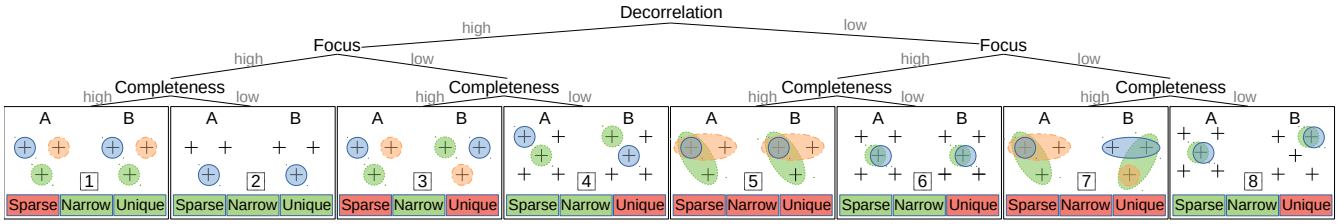


Figure 2: **Relation between SPARROW measures (tree nodes) and explainability principles (leafs).** Each leaf contains parts (crosses) of two schematic samples (A and B), where the parts in A and B are the same at the respective positions. Overlayed are schematic prototype masks (blue, green, orange) which depict a typical situation corresponding to the specific value scale (high, low) of SPARROW measure scores. The bottom of each leaf displays the leaf number and the consequence for each principle (green: fulfilled, red: not fulfilled).

The decorrelation of prototypes can be determined by first counting per sample the number of parts that are matched by u prototypes. We call this count $\text{cnt}^{\text{decorr}}_{i,u}$, where $u \in [1,\ldots,|\mathcal{P}_{c_i}|]$ with the number of prototypes $|\mathcal{P}_{c_i}|$ in class c_i of the sample with index i. In figure 2, leaf 5, we would have $\text{cnt}^{\text{decorr}}_{\text{idx}(A),1} = 2$, $\text{cnt}^{\text{decorr}}_{\text{idx}(A),2} = 0$ and $\text{cnt}^{\text{decorr}}_{\text{idx}(A),3} = 1$. Perfect decorrelation would require $\text{cnt}^{\text{decorr}}_{i,u>1} = 0$ for every sample index i. In addition, the higher the number of prototypes which match the same part becomes the more correlated this set of prototypes becomes. Motivated by this idea we define the first measure.

Definition 1 (Prototype decorrelation) *The prototype decorrelation (ptd) is defined as the normalized weighted sum over the number of times that annotated parts are matched by prototype activation masks over all samples:*

$$\text{ptd} = \sum_{i,u} \frac{(|\mathcal{P}_{c_i}| + 1 - u) \cdot \text{cnt}^{\text{decorr}}_{i,u}}{N \cdot \tilde{T}_i \cdot |\mathcal{P}_{c_i}|}, \tag{3}$$

where N is the total number of samples and $\tilde{T}_i$ is the number of parts for sample i which are matched by at least one prototype ($\tilde{T}_i > 0$ as discussed before). $\text{ptd}_{\max} = 1.0$ is achieved when no part is matched by more than one prototype. Then, $\text{cnt}^{\text{decorr}}_{i,u=1} = \tilde{T}_i$, and $\text{cnt}^{\text{decorr}}_{i,u>1} = 0$ for all i. On the other hand, $\text{ptd}_{\min}$ arises when all parts are matched by all prototypes for every sample. Then, $cnt_{i,u=|P_{c_i}|} = \tilde{T}_i$ and $cnt_{i,u\neq|P_{c_i}|} = 0$ for all i. If we neglect pruning, i.e. $|\mathcal{P}_{c_i}| = \frac{m}{C}$ with the number of prototypes m and classes C for all c_i, we arrive at $\text{ptd}_{\min} = \frac{C}{m}$. Since it only makes sense to look into the decorrelation of prototypes per class for $m > C$ (where typically $m \gg C$), $\text{ptd}_{\min}$ will typically be a small number. This result also means that for only one prototype per class *ptd* yields a perfect decorrelation score which would be expected since it is defined per class.

The prototype focus level follows the goal of determining how consistent prototypes are in recurrently matching the same part over all samples. A prototype that matches the head of a bird in one picture but the tail in another would not be well focused. Prototypes with little focus either consistently represent multiple part semantics (*e.g.* always the breast and the belly of a bird) or represent semantics that are not purely part-related (they might *e.g.* focus on color or texture) or a mixture of both. These cases will be further discussed in section 4.

For each prototype and all T types of annotated parts we count the number of times that this prototype globally (i.e. over all samples) matches this part type. We normalize this

quantity by dividing each count by the total number of matches between this prototype and any part in any sample. We call it $\text{frac}_{j,k}^{\text{focus}}$, where $j \in [1,\dots,m]$ and $k \in [1,\dots,T]$ with the total number of prototypes m and parts T. $\sum_k \text{frac}_{j,k}^{\text{focus}} = 1$ for each prototype with index j due to normalization. A highly focused prototype would have very sparse values in $\text{frac}_{j,k}^{\text{focus}}$. In order to quantify this concept over all prototypes we suggest looking at the top-1 matched part for each prototype which leads to the distribution $\text{frac}_{j}^{\text{focus-top-1}} = \max_k(\text{frac}_{j,k}^{\text{focus}})$. We derive the following measure:

Definition 2 (Prototype focus) *The prototype focus (ptf) is defined as the median of the distribution of normalized global match counts of prototypes with their top-1-matched part:*

$$\text{ptf} = \text{median}_j(\text{frac}_{j}^{\text{focus-top-1}}). \tag{4}$$

For samples A and B in figure 2, leaf 7, we can find $\text{frac}_{\text{idx(blue)}}^{\text{focus-top-1}} = \frac{2}{3}$, $\text{frac}_{\text{idx(green)}}^{\text{focus-top-1}} = \frac{1}{2}$ and $\text{frac}_{\text{idx(orange)}}^{\text{focus-top-1}} = \frac{1}{3}$, so that $\text{ptf} = \frac{1}{2}$. For leaf 1 instead, we find $\text{ptf} = 1$.

The completeness of the description of a sample by prototypes is the last concept that we present. It quantifies how fully samples are described by prototypes in terms of the annotated parts. A sample would be completely described by prototypes if all its parts were matched by at least one prototype. This is a useful concept to track in order to balance the trade-off between the sparsity of explanations and a full description of samples by annotated part semantics that are deemed important by domain experts. This trade-off exists since it is easiest to maximize completeness by increasing the number of prototypes. This point is discussed in more detail at the end of this section.

To derive the completeness measure, we start by counting how many parts are matched by prototypes per sample. We name this count $\text{cnt}_i^{\text{comp}}$ and it is $\text{cnt}_i^{\text{comp}} \in [1,\dots,T]$ for each sample with index i and with the total number of parts T. For example in figure 2, leaf 4, we have $\text{cnt}_{\text{idx}(A)}^{\text{comp}} = 2$. Since it is clear that samples with lower numbers of captured parts by prototypes are worse with respect to the completeness of the sample description we can define the completeness measure as follows.

Definition 3 (Completeness of sample description) *The completeness of sample description (sac) is defined as the normalized sum over the number of annotated parts which are matched by at least one prototype activation mask over all samples:*

$$\text{sac} = \sum_i \frac{\text{cnt}_i^{\text{comp}}}{N \cdot T}. \tag{5}$$

For the samples A and B in leaf 4 (figure 2) this yields $\text{sac} = \frac{2+2}{2 \cdot 5} = 0.4$. We see that $\text{sac}_{\text{max}} = 1.0$ which requires all parts in all samples to be captured by prototypes. In practice this may not be possible if not all parts are visible for every sample and in this case are not annotated. The theoretical minimum score is $\text{sac}_{\text{min}} = \frac{1}{T}$ which would happen if every sample had exactly one part matched by all prototypes ($\text{cnt}_i^{\text{comp}} = 1$ for all i). If it is the goal to maximize both *ptd* and *sac* measures, we suggest to use the following overall measure.

Definition 4 (Decorrelation-completeness balance) *The decorrelation-completeness balance (dcb) between the ptd and sac measures is defined as their harmonic mean as:*

$$\text{dcb} = 2 \cdot \frac{\text{ptd} \cdot \text{sac}}{\text{ptd} + \text{sac}}. \tag{6}$$

Take note that in order to achieve a perfect *dcb* score it is necessary that the number of prototypes per class is exactly as high as the number of annotated parts per sample, which means that (without pruning) the total number of prototypes must be $m = C \cdot T$. Otherwise, since a prototype is always matched with at least one part, a perfect *sac* score could only be achieved by a reduced *ptd* score.

Finally, we look at the relation between the SPARROW measures and the explainability principles introduced in section 1. We can see at the bottom of figure 2 that the only way to fulfill all principles is to achieve high *ptd* and *ptf* scores and a low *sac* score (leaf 2) which would require to have only one prototype per class. One might argue that leaf 1 is actually preferable, i.e. favoring completeness over sparsity of explanations. There are valid arguments for this choice. For once, it is generally known that there is a trade-off between sparsity and classification performance [21]. Additionally, in the computer vision domain it is not believed that fewer pixels generally constitute a better explanation [21]. However, in the case of explanations by prototypes we argue that the similarity of a sample to prototypes has to be verified by the end user separately for each prototype. Cognitive science indicates that people rarely expect complete explanations but are typically content with a few presented causes for a decision [15]. Grasping many explanations at the same time is difficult for humans since their mental capacity is limited to process 7 ± 2 items at once [14]. In [18] predictions are explained with hierarchical prototype trees and the authors share the belief that smaller trees with fewer prototypes are easier to interpret. In the end - although we speculate that sparse explanations may be preferable and one will typically want to achieve a situation in between those depicted in leaf 1 and leaf 2 - we follow the assessment of [16] in that the importance of sparsity should be measured by empirical evaluation which we plan to perform in the future.

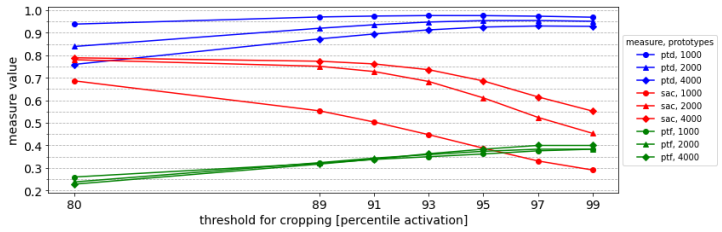
3 Experiments

In the following, “ProtoPNet" denotes the model described in section 1. “SPARROW (own)" includes the additional novel loss components discussed in section 2. For additional details about the experiments, *e.g.* the dataset, data preprocessing, hyperparameters, training hardware or time demand we refer to the supplementary material.

Dataset The subsequent experiments are performed on the Caltech-UCSD Birds-200-2011 dataset (CUB) [23]. This dataset is selected since it allows for a direct comparison to the results from Chen *et al.* [6].

Hyperparameter Tuning We tuned the loss component weights by a combination of random search and Bayesian Optimization to the values reported in table 1.

Another important hyperparameter is the choice of the threshold for activation masks (cf. section 2). This choice is driven by the preference if only the most salient similarities should be kept (high threshold) or if also minor similarities should be taken into account (low threshold). We followed Chen *et al.* [6], who selected the 95-th percentile, in opting for a high threshold. However, we wanted to analyze the effect that varying the threshold in the high value range has on the SPARROW measures. This is shown in figure 3 (a). We can see that *sac* is most sensitive to the threshold whereas *ptd* and *ptf* vary to a lesser extent. We recommend to chose the 95-th percentile when comparing a method to other works but we emphasise that other choices are valid as well.

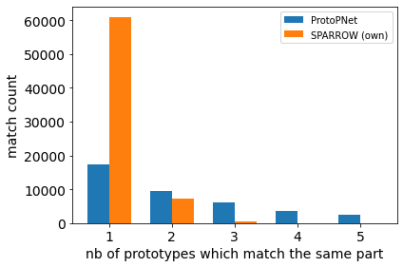


(a)

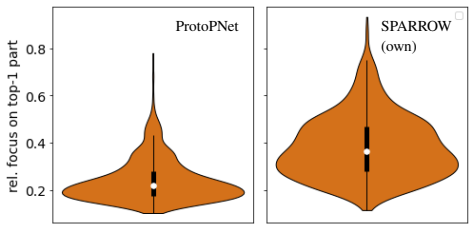
ProtoPNet		Own	
Prototypes	Accuracy	Prototypes	Accuracy
1000	**73.4** $\pm$ **0.3**	1000	70.3 $\pm$ 1.1
2000	73.3 $\pm$ 1.5	2000	**74.7** $\pm$ **0.7**
4000	73.7 $\pm$ 1.4	4000	**75.9** $\pm$ **1.4**

(b)

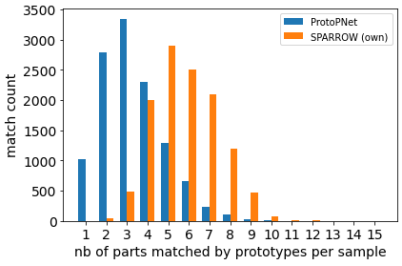
Figure 3: **(a) Dependency of SPARROW measures on the cropping threshold** for activation masks. **(b) Comparison of the average accuracy** of ProtoPNet with our own method.



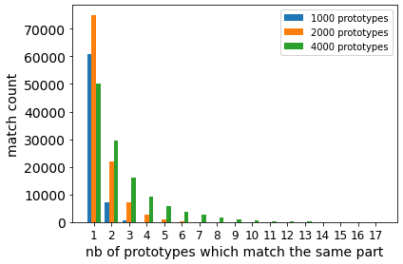
(a) Distributions leading to *ptd* measures for two models



(b) Distributions leading to *ptf* measures for two models



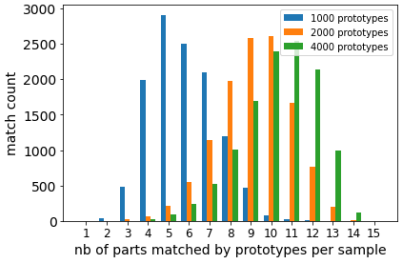
(c) Distributions leading to *sac* measures for two models



(d) Distrs. leading to *ptd* measures the SPARROW model

Model	Prototypes	Measures			
		ptd	*sac*	*dcb*	*ptf*
ProtoPNet	1000	78.4	22.1	34.5	21.7
SPARROW (own)	1000	**97.6**	38.8	55.5	36.2
SPARROW (own)	2000	95.3	61.1	74.5	37.4
SPARROW (own)	4000	92.5	**68.6**	**78.8**	**38.3**

(e) SPARROW evaluation measures



(f) Distrs. leading to *sac* measures for the SPARROW model

Figure 4: **Comparing the explanatory capacity of prototypes.** Top row: Distributions leading to the interpretability measures for 1000 prototypes and different models (from left to right): Prototype decorrelation (*ptd*), prototype focus (*ptf*) and completeness of sample description (*sac*). Bottom row: Left and right report results for the SPARROW model for various numbers of prototypes. Results for the SPARROW measures are shown in the table.

Classification Performance We compare the classification performance of our method with ProtoPNet as shown in figure 3 (b). The performance of the two models is overall comparable. ProtoPNet shows better performance for a low number of prototypes while our SPARROW method shows the potential to perform better for higher numbers of prototypes.

Evaluation with SPARROW We start by comparing ProtoPNet with our method, where both methods use 1000 prototypes. The table in figure 4 (e) shows that there is a significant improvement in all four measures. This implies that prototypes produced by our method are at the same time more decorrelated from each other, more focused towards single semantic concepts while representing more semantic concepts in samples than ProtoPNet. This is reflected in the distributions from which the measures are derived. There are significantly more prototypes that match only one part per sample and significantly fewer prototypes overlapping at high numbers of same-sample matches (cf. figure 4 (a)) which leads to the higher *ptd* score. The distribution of our run in figure 4 (b) is shifted towards the perfect score of 1.0 when compared with ProtoPNet. This leads to the higher *ptf* score which is the

median (white dot) of the distributions. Finally the distribution in figure 4 (c) for our method is shifted to the right when compared with ProtoPNet towards the theoretical goal of the *sac* measure of all counts being at mark 15. This comes as a bit of a surprise since our method was not designed to increase the completeness of the sample description. For just 1000 prototypes (i.e. 5 prototypes per class) it is actually more useful to look at this distribution instead of the *dcb* measure since a perfect *dcb* score is not possible (cf. discussion after definition 4). We can see that the maximum of the distribution is at mark 5 which seems ideal since counts at marks < 5 indicate an imperfect decorrelation and counts at marks > 5 an imperfect focus of the prototypes.

Up next, we compare the models based on our method for different numbers of prototypes. The distributions (d) and (f) again show additional details for the calculation of the measures. From the table in figure 4 (e) we see that an increased number of prototypes leads to a decreased decorrelation (*ptd*) and an increased completeness of sample description (*sac*) which we would intuitively have expected. This shows that it made sense to define the decorrelation-completeness balance measure (*dcb*) as the harmonic mean between *ptd* and *sac*. Based on *dcb* alone, 4000 prototypes seem optimal. However, if *ptd* or sparsity of explanations is considered most important, a lower number of prototypes might be preferable (cf. the discussion about sparsity at the end of section 2). Looking at the *ptf* measure we see an increased focus with an increasing number of prototypes. 1000 prototypes may not have been enough to capture all important semantics for optimizing the classification score so that the focus of prototypes was more “washed out" to compensate for this.

We can conclude that the quantification of the explanatory capacity of prototypes by means of the SPARROW evaluation protocol confirms the improved visual interpretability of prototypes from our method when compared with ProtoPNet (cf. case study in section 1).

4 Scope and Future Extensions

The SPARROW measures are guided by well-known principles of interpretability (cf. section 1). They are derived from latent space activations of a CNN which are then used as prototypes in the prediction following the ProtoPNet method [6]. Since ProtoPNet is model agnostic in the sense that it allows to use arbitrary convolutional base networks so is our evaluation protocol. A limitation that remains is the reliance on part semantics in the form of keypoint annotations. SPARROW is therefore best used if there is domain-specific indication that an image classification dataset contains part-related concepts which are well suited for a corresponding prediction task. SPARROW then allows to optimize models to contain semantically coherent prototypes representing those concepts. Such models are expected to be useful for human experts like ornithologists or physicians for whom it is a common strategy to explain class predictions based on part-related semantics [6] in a case-based reasoning fashion [13, 21]. Examples for suitable tasks and datasets which already contain keypoint part annotations are the prediction of animal species like birds [23] or tigers [11]. Unfortunately not many datasets currently contain ground truth annotations at the part level. Oftentimes there is no way around employing domain experts to perform the labeling. This issue is not just symptomatic of our proposed evaluation protocol but is frequently found in the area of evaluating conceptual representations and *e.g.* lead [4] to release the Broden dataset which contains diverse conceptual annotations.

For other tasks like human pose estimation [3, 8] concepts that relate to a combination of parts (*e.g.* the relative position of joints to each other) are deemed important [5]. SPARROW

could still be applied to such tasks but it would make sense to relax the definition of the prototype focus measure to tolerate prototypes which recurrently focus on the same group of parts instead of single parts. Apart from keypoint part annotations many datasets also use other types of concept annotations like pixel-wise binary masks or bounding boxes as well as attributes or relations between those annotations which are especially useful to label higher level concepts [4, 9]. In order to match similarity maps of samples and prototypes with pixel-wise annotations, the approach in [4] can be used. Also, harder tasks like semantic image retrieval are generally expected to require models to learn higher level semantic concepts [7, 12]. We therefore want to evolve SPARROW to be applicable to different types of possibly spatially or semantically overlapping concept annotations on different levels of abstraction. This would require to adapt the prototype focus measure. However, if *e.g.* bounding boxes are restricted around relatively small image parts, they could be converted to keypoint part annotations and it would make sense to apply SPARROW in its current form.

An interesting extension would be to incorporate the importance of visual characteristics (*e.g.* texture, shape, hue) about part-related concepts that prototypes represent into SPARROW. This knowledge can be obtained without additional human annotations as was shown by [17]. Also pruning of prototypes which focus on background semantics [6] should be helpful because such prototypes do not generalize well and distort the results of SPARROW.

Currently, the SPARROW evaluation protocol encourages single prototypes per part per class. An alternative would be to enable sharing prototypes between classes. This could potentially lead to a better global interpretability. We might for example find inter-class prototypes for short and long bird legs which could be useful to discriminate between different bird species. We believe that it comes down to the goal (*e.g.* knowledge discovery or decision support) which approach is preferable. For example end users might be confused if a sample from which a prototype originates has a different class identity than test samples that are explained by being similar to this prototype. This might diminish trust in the system.

Finally, we plan to break down the global analysis from SPARROW to the local prototype level. This way, SPARROW could be used to find prototype candidates for pruning. Additionally it could be used in concept discovery in an expert-in-the-loop approach to annotate new concepts and evaluate prototypes iteratively. For this purpose, prototype activations in samples that do not match any annotated concept should be investigated closely instead of automatically selecting the nearest part as a match, as it is currently done. In this respect the location instability measure [24] should prove useful to find new part-related concepts – something that the prototype focus measure cannot accomplish.

5 Conclusion and Outlook

The rapidly growing number of available prototype classification methods calls for standardized and efficient ways to assure the quality of a new technique in comparison with other approaches on various datasets. Quality assurance is a key aspect of explainable models since those explanations determine how well humans can understand the model predictions. In this work, we presented SPARROW, an own prototype generation method and a benchmarking protocol for the standardized and transparent comparison of prototype based classification methods. In the explainability field, SPARROW bears the potential to help researchers and practitioners alike to efficiently derive more realistic and use-case-driven prototype models and assure their quality through extensive comparative evaluations. We hope that this work contributes to further advances in explainability research.

Acknowledgements

We thank Mr. Martin Pawelczyk for many helpful discussions and comments. We also thank Ms. Elizabeth Baker for proofreading.

References

[1] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. In *Advances in Neural Information Processing Systems (NeurIPS) 31*.

[2] David Alvarez-Melis and Tommi S Jaakkola. Towards robust interpretability with self-explaining neural networks. In *Advances in Neural Information Processing Systems (NeurIPS) 31*, 2018.

[3] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *Proceedings of the IEEE Conference on computer Vision and Pattern Recognition*, pages 3686–3693, 2014.

[4] David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. Network dissection: Quantifying interpretability of deep visual representations. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2017.

[5] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7291–7299, 2017.

[6] Chaofan Chen, Oscar Li, Daniel Tao, Alina Barnett, Cynthia Rudin, and Jonathan K Su. This looks like that: deep learning for interpretable image recognition. In *Advances in neural information processing systems (NeurIPS)*, 2019.

[7] Justin Johnson, Ranjay Krishna, Michael Stark, Li-Jia Li, David Shamma, Michael Bernstein, and Li Fei-Fei. Image retrieval using scene graphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3668–3678, 2015.

[8] Sam Johnson and Mark Everingham. Clustered pose and nonlinear appearance models for human pose estimation. In *bmvc*, volume 2, page 5. Citeseer, 2010.

[9] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73, 2017.

[10] Oscar Li, Hao Liu, Chaofan Chen, and Cynthia Rudin. Deep learning for case-based reasoning through prototypes: A neural network that explains its predictions. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2018.

[11] Shuyuan Li, Jianguo Li, Hanlin Tang, Rui Qian, and Weiyao Lin. Atrw: A benchmark for amur tiger re-identification in the wild. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 2590–2598, 2020.

[12] Cewu Lu, Ranjay Krishna, Michael Bernstein, and Li Fei-Fei. Visual relationship detection with language priors. In *European conference on computer vision*, pages 852–869. Springer, 2016.

[13] Cindy Marling, Stefania Montani, Isabelle Bichindaritz, and Peter Funk. Synergistic case-based reasoning in medical domains. *Expert systems with applications*, 41(2): 249–259, 2014.

[14] George A Miller. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological review*, 63(2):81, 1956.

[15] Tim Miller, Piers Howe, and Liz Sonenberg. Explainable ai: Beware of inmates running the asylum or: How i learnt to stop worrying and love the social and behavioural sciences. *arXiv preprint arXiv:1712.00547*, 2017.

[16] Menaka Narayanan, Emily Chen, Jeffrey He, Been Kim, Sam Gershman, and Finale Doshi-Velez. How do humans understand explanations from machine learning systems? an evaluation of the human-interpretability of explanation. *arXiv preprint arXiv:1802.00682*, 2018.

[17] Meike Nauta, Annemarie Jutte, Jesper Provoost, and Christin Seifert. This looks like that, because... explaining prototypes for interpretable image recognition. *arXiv preprint arXiv:2011.02863*, 2020.

[18] Meike Nauta, Ron van Bree, and Christin Seifert. Neural prototype trees for interpretable fine-grained image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14933–14943, 2021.

[19] Hieu V Nguyen and Li Bai. Cosine similarity metric learning for face verification. In *Asian conference on computer vision*, pages 709–720. Springer, 2010.

[20] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 2019.

[21] Cynthia Rudin, Chaofan Chen, Zhi Chen, Haiyang Huang, Lesia Semenova, and Chudi Zhong. Interpretable machine learning: Fundamental principles and 10 grand challenges. *arXiv preprint arXiv:2103.11251*, 2021.

[22] Anshumali Shrivastava and Ping Li. Improved asymmetric locality sensitive hashing (alsh) for maximum inner product search (mips). In *31st Conference on Uncertainty in Artificial Intelligence, UAI 2015*, 2015.

[23] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The Caltech-UCSD Birds-200-2011 Dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011.

[24] Quanshi Zhang, Ruiming Cao, Feng Shi, Ying Nian Wu, and Song-Chun Zhu. Interpreting cnn knowledge via an explanatory graph. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[25] Quanshi Zhang, Ying Nian Wu, and Song-Chun Zhu. Interpretable convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

B Supplementary Material for SPARROW: Semantically Coherent Prototypes for Image Classification

Supplementary material of the following publication is included in this appendix:

Stefan Kraft, Klaus Broelemann, Andreas Theissler, and Gjergji Kasneci. SPARROW: Semantically Coherent Prototypes for Image Classification. In *BMVC*, page 186, 2021

Supplementary Material for SPARROW: Semantically Coherent Prototypes for Image Classification

Stefan Kraft[1,4]
stefan.kraft@stz-softwaretechnik.de
Klaus Broelemann[2]
klaus.broelemann@schufa.de
Andreas Theissler[3]
https://orcid.org/0000-0003-0746-0424
Gjergji Kasneci[4,2]
gjergji.kasneci@uni-tuebingen.de

[1] IT-Designers Group
Esslingen am Neckar, GER
[2] SCHUFA Holding AG
Wiesbaden, GER
[3] Aalen University of Applied Sciences
Aalen, GER
[4] Data Science & Analytics Research
The University of Tübingen
Tübingen, GER

This supplementary material is structured as follows. We start by summarizing the mathematical notation in section 1 and proceed to give additional details about the model architecture (section 2), training process (section 3), `SPARROW` framework (section 4) and experiments (section 5) presented in the main paper.

1 Notation

Table 1 summarizes the conventions regarding the notation and meaning of mathematical symbols in the main paper and the remainder of this supplementary material.

2 Model architecture

This section provides additional details about the model architecture of ProtoPNet [1]. An overview is shown in table 2. The prototype classification network learns a transformation $h \circ g_P \circ f(X)$ on images from a given dataset $D = [X,Y] = \{\mathbf{x}_i, y_i\}_{i=1}^n$. The transformation is composed of a convolutional base network f, a prototype unit g_P and a fully-connected layer h.

Convolutional base network The convolutional base network f learns a latent space feature embedding. We select a ResNet18 architecture to which we append two convolutional addon-layers which do not change the shape of the ResNet18 output in accordance with Chen *et al.* [1]. For the ResNet18 layers we use weights pretrained on ImageNet [5]. f reduces the sample dimensions from S_{raw} to $S = (s,s) = (7,7)$ and the whole latent space dimension is $(l,s,s) = (512,7,7)$, where l is the number of convolutional filter layers.

Symbol	Meaning
n_{train}	Number of unaugmented training samples
n_{test}	Number of test samples
N	Total number of unaugmented samples in the dataset. $N = n_{\text{train}} + n_{\text{test}}$
n_{aug}	Number of augmented samples
$\rho_{\text{train-val}}$	Fraction of augmented samples n_{aug} which are used for training
n	Number of augmented training samples, given by $n = \rho_{\text{train-val}} \cdot n_{\text{aug}}$
n_{val}	Number of augmented validation samples, given by $n_{\text{val}} = (1 - \rho_{\text{train-val}}) \cdot n_{\text{aug}}$
S_{raw}	Size of samples after preprocessing (cropping and rescaling)
S	Kernel size of samples in latent space which is quadratic: $S := (s, s)$
S_{pt}	Kernel size of prototypes in latent space which is quadratic: $S_{\text{pt}} := (s_{\text{pt}}, s_{\text{pt}})$
l	Number of convolutional filter layers in latent space
m	Number of prototypes before pruning
C	Number of classes in the dataset
$\mathbf{x}_i$	Single preprocessed training sample, $i \in [1, \dots, n]$
c_i	Class of sample $\mathbf{x}_i$
X	Set of all preprocessed training samples$\{\mathbf{x}_{i=1}^{n}\}$
$\mathbf{p}_j$	Single prototype, $j \in [1, \dots, m]$
P	Set of all prototypes $\{\mathbf{p}_{j=1}^{m}\}$
P_{c_i}	Set of prototypes $\{\mathbf{p}_{j=\tilde{m}_i}^{\tilde{m}_i\prime}\}$ of class c_i with $\tilde{m}_i, \tilde{m}_i\prime \in [1, \dots m]$
$\mathcal{P}_{c_i}$	Interval of indices $[\tilde{m}_i, \dots, \tilde{m}_i\prime]$ for prototypes of class c_i with $\tilde{m}_i, \tilde{m}_i\prime \in [1, \dots m]$
f	Combination of base- and add-on layers of the neural network
g_P	Prototype unit of the neural network which contains prototypes P
h	Last layer of the neural network
$\mathcal{L}$	Total loss in the prototype learning phase of the training
$\mathcal{L}^{\nu}$	ν-th loss component of $\mathcal{L}$ with $\nu \in \{\text{CrsEnt}, \text{Clst}, \text{Sep}, \text{AS}, \text{PSD}\}$
$\lambda_{\mathcal{L}^{\nu}}$	Weight of the ν-th loss component $\mathcal{L}^{\nu}$ in the total loss $\mathcal{L}$
$\omega_h^{\gamma,j}$	Last layer weights with $\gamma \in [1, \dots, C]$ and $j \in [1, \dots m]$
$\mathcal{L}^{L1}$	$L1$ norm regularization of last layer weights $\omega_h^{\gamma,j}$
$\lambda_{\mathcal{L}^{L1}}$	Weight of the regularizer $\mathcal{L}^{L1}$ in the total last layer optimization loss
T	Number of annotated part types in the dataset
$\tilde{T}_i$	Number of parts in the i-th sample which are matched by at least one prototype
$\mathbf{r}_{i,k}$	Coordinate of the k-th part in the transformed sample input space of the i-th sample
R	Set of part coordinates $\{\mathbf{r}_{i,k=1,1}^{N,T}\}$
$\text{match}_{i,j,k}$	Scalar which is 1 if there is a match between the j-th prototype ($j \in \mathcal{P}_{c_i}$) and the k-th part ($k \in [1, \dots, T]$) for the i-th sample ($i \in [1, \dots, N]$). Otherwise it is 0

Table 1: **Table of notation**

Hierarchy	Type	Output size	Info
Base	ResNet	$(512,7,7)$	ResNet architecture [3], based on the torchvision implementation [5], where we stripped away the last two layers (average pooling and fully connected layer)
Add-on	conv	$(512,7,7)$	Kernel size $(1,1)$, stride $(1,1)$, ReLU activation function, Kaiming He initialization [2]
Add-on	conv	$(512,7,7)$	Kernel size $(1,1)$, stride $(1,1)$, Sigmoid activation function, Kaiming He initialization [2]
Prototype	params	$(512,s_{\text{pt}},s_{\text{pt}})$	Kernel size $(s_{\text{pt}},s_{\text{pt}})$, parameters are randomly initialized from a uniform distribution on $[0,1)$
Last-layer	fc	(m,C)	No bias, Kaiming He initialization [2]

Table 2: **Model architecture** of ProtoPNet

Prototype unit The prototype unit g_P contains a set of m prototypes $P = \{\mathbf{p}_{j=1}^{m}\}$ which occupy a latent space kernel of the minimal possible size $S_{pt} = (s_{pt}, s_{pt}) = (1,1)$ and therefore have the shape (m,l,s_{pt},s_{pt}). For each sample, g_P first calculates the squared $L2$ distances between the latent space patches of samples $f(X)$ and the prototypes P. This calculation is performed per prototype while summing over the filter layers. The resulting distances are converted to similarity scores. Concretely, this means that the j-th prototype unit $g_{\mathbf{p}_j}$ computes a similarity score between prototype $\mathbf{p}_j$ and latent space patches of a sample $\xi \in X$ as [1]

$$g_{\mathbf{p}_j}(f(\xi)) = \max_{\mathbf{z} \in patches(f(\xi))} \log\left(\frac{\|\mathbf{z}-\mathbf{p}_j\|^2+1}{\|\mathbf{z}-\mathbf{p}_j\|^2+\varepsilon}\right), \tag{1}$$

where ε is a very small number. Since this is a monotonically decreasing function with respect to the $L2$ distance $\|\mathbf{z}-\mathbf{p}_j\|$ between a prototype and a latent space patch of a sample, it is a suitable choice for describing the similarity between a prototype and sample patch [1]. This means that the similarity between a sample and a prototype is effectively determined by the latent space patch of the sample which is most similar to the given prototype. The similarity scores have the shape (m,s_{pt},s_{pt}) which in our case is $(m,1,1)$.

Fully connected layer The fully connected layer h (called “last layer") calculates final class scores as a weighted superposition of the output of g_P. This is done by applying a fully connected layer without a bias term on the resulting similarity scores which is of shape (m,C) where C is the number of classes.

3 Training process

In this section we provide further details about the training process of ProtoPNet [1].

Training steps The training consists of three main steps. As a final step prototypes that focus on background semantics may be pruned, *e.g.* with the approach by Chen *et al.* [1]. In this work we do not perform any pruning.

Step 1: Prototype learning Parameters of the neural network and the prototypes are jointly learned while the weights of the last layer are frozen. An important concept is that the weights of the last layer are fixed such that each output logit is connected to an equal

number of outputs of g_P while all other connections are set to -0.5. For this reason it can be said that prototypes “have a certain class identity" or “belong to a class". The total loss function is composed of a weighted superposition of five loss components with static weights as (cf. table 1 in the main paper): $\mathcal{L} = \sum_\nu \lambda_{\mathcal{L}^\nu} \cdot \mathcal{L}^\nu$ with $\nu \in \{\text{CrsEnt}, \text{Clst}, \text{Sep}, \text{AS}, \text{PSD}\}$. Only class labels but no part annotations are used in training.

Step 2: Prototype projection After the main training is finished a prototype projection (also called “prototype push") is performed by which the prototypes are moved in latent space to the sample patch with the same class identity as the prototype which has the minimum $L2$-distance of all samples in the training set.

Step 3: Last layer optimization The prototype push is followed by a convex optimization of the last layer where the weights of all other layers are frozen. During this phase a $L1$ norm regularization of the off-label weight connections $\omega_h^{\gamma,j}$ between the C classes ($\gamma \in [1,\dots,C]$) and m outputs ($j \in [1,\dots m]$) of the prototype unit g_P is performed as

$$\mathcal{L}^{L1} = \sum_\gamma \sum_{j \notin \mathcal{P}_\gamma} |\omega_h^{\gamma,j}|, \tag{2}$$

where $\mathcal{P}_\gamma$ is the set of prototypes of the γ-th class. The off-label weights which were initially fixed to -0.5 are effectively reduced to zero. This is done to prevent a negative reasoning process for the classification [1]. This loss is added to the total loss (which at this stage is given by $\mathcal{L}^{\text{CrsEnt}}$) with weight $\lambda_{\mathcal{L}^{L1}}$. The last layer optimization is performed over $\text{epochs}_{\text{ll}}$ epochs.

Train-validation split Before the training, we perform a train-validation split where we split the n_{aug} augmented samples into $n = (\rho_{\text{train-val}}) \cdot n_{\text{aug}}$ training and $n_{\text{val}} = (1 - \rho_{\text{train-val}}) \cdot n_{\text{aug}}$ validation samples, where $\rho_{\text{train-val}}$ is the splitting factor.

Warm start The training starts with a “warm start" of several epochs in which only the parameters of the add-on layers and the prototype vectors P are optimized while the parameters of the base network and the last layer are frozen. Only prototype learning is performed during the warm start. The training phase after the warm start is refered to as “main training".

Early stopping Both the warm start and main training run until stopped by an early-stopping method when the average validation accuracy does not increase by the fraction $\text{delta}_{\text{warm/main}}$ of the validation score over $\text{pat}_{\text{warm/main}}$ epochs (patience). After early stopping of the prototype learning step in the main training phase, the model checkpoint of pat_{main} epochs earlier is loaded and used for the following steps, i.e. prototype push and last layer optimization.

Clamping of cosine similarity The cosine similarity $\text{CS}(\mathbf{p}_i, \mathbf{p}_j) = \frac{\mathbf{p}_i^T \cdot \mathbf{p}_j}{\|\mathbf{p}_i\| \|\mathbf{p}_j\|}$ between two prototypes $\mathbf{p}_i$ and $\mathbf{p}_j$ can theoretically take its maximum or minimum value of 1 and -1. Especially the former will occasionally happen after a prototype push if two prototypes are pushed to the same latent space sample patch of the same sample. This is not a problem in principal [1] but will lead to infinite gradients of $\text{AS}(\mathbf{p}_i, \mathbf{p}_j) = 1 - \frac{1}{\pi} \arccos(\text{CS}(\mathbf{p}_i, \mathbf{p}_j))$ during backpropagation. To prevent this we decided to clamp the values of the cosine similarity to the interval $[-0.9999, 0.9999]$.

Simplifications In order to simplify the training process and since we are not interested in optimizing the classification performance of the baseline in the context of this work we only perform one final push epoch and also do not use a learning rate scheduler in contrast to Chen *et al.* [1].

4 SPARROW

Matching Prototypes with Part Annotations This section gives a more detailed explanation on how to obtain the matches between prototypes and part annotations (cf. section 2 of the main paper). We define that the part annotations $R = \{\mathbf{r}_{i,k=1,1}^{N,T}\}$ have already been subject to the same transformations as the raw input samples (cf. section 5). As noted in algorithm 1 we first obtain the activation map of the latent space sample patch which has the smallest $L2$ distance to a prototype (line 5). This is done for all samples $\{\mathbf{x}_{i=1}^{N}\}$ and prototypes $P_{c_i} = \{\mathbf{p}_{j=\tilde{m}_i}^{\tilde{m}_i\prime}\}$ with the same class identity $c_i \in [1,\dots,C]$ as the samples. Here, $\tilde{m}_i, \tilde{m}_i\prime \in [1,\dots m]$ and we define $\mathcal{P}_{c_i} = [\tilde{m}_i,\dots,\tilde{m}_i\prime]$ to be the interval of indices of prototypes which belong to class c_i. The activation maps are upsampled to the pixel space of sample inputs (line 6). Next, a mask is selected based on an activation threshold (line 7).

Schematic prototype masks are illustrated together with annotated image parts in figure 1. Each part coordinate is matched with the mask and matches are noted in $\{match_{i,j,k=1,\tilde{m}_1,1}^{N,\tilde{m}_N,T}\}$ (lines 8-13). If an activation mask does not match any part annotation, we select the closest coordinate to the mask as a match (lines 14-17).

Algorithm 1 Find matches between sample patch activations by prototypes and part annotations.

1: $\text{match}_{i,j,k} \leftarrow$ initialize to 0 for $(i,j,k) = (1,\tilde{m}_1,1),\dots,(N,\tilde{m}_N,T)$
2: **for** $i = 1,2,\dots,N$ **do**
3: **for** $j = \tilde{m}_i\dots,\tilde{m}_i'$ **do**
4: pt_has_match$\leftarrow$ *False*
5: $\text{act}_{i,j} \leftarrow \min_{\mathbf{z}\in patches(f(\mathbf{x}_i))} \|\mathbf{z}-\mathbf{p}_j\|^2$
6: $\text{act}_{i,j}^{up} \leftarrow$ upsampling of $\text{act}_{i,j}$ to pixel space
7: $\text{mask}_{i,j} \leftarrow$ select values $>$ threshold in $\text{act}_{i,j}^{up}$
8: **for** $k = 1,2,\dots,T$ **do**
9: **if** $\mathbf{r}_{i,k}$ lies within the mask **then**
10: $\text{match}_{i,j,k} \leftarrow 1$
11: pt_has_match$\leftarrow$ *True*
12: **end if**
13: **end for**
14: **if** not pt_has_match **then**
15: Find k for which $\mathbf{r}_{i,k}$ is closest to the mask
16: $\text{match}_{i,j,k} \leftarrow 1$
17: **end if**
18: **end for**
19: **end for**

Figure 1: **Part annotations vs. activation masks.** Sample images from the CUB [6] dataset (containing $T = 15$ part annotations) showing annotations of visible parts (numbers 1 to 15) and schematic illustrations of the activation masks of prototypes (P1 to P5).

5 Experiments

This section provides additional details about the experiments that we performed (cf. section 3 of the main paper).

Dataset The Caltech-UCSD Birds-200-2011 dataset (CUB) [6] contains $N = 11788$ images of birds which divide into $n_{\text{train}} = 5994$ training and $n_{\text{test}} = 5794$ test samples. For each sample there exists one class- and $T = 15$ keypoint part annotations[1], where the total number of classes is $C = 200$ bird species. The samples are distributed reasonably evenly between classes so that one class contains about 30 samples. After preprocessing which includes data augmentation we have $n = 194206$ augmented training samples of size $S_{\text{raw}} = (224, 224)$ and each class contains about 1200 images. Training is performed on the augmented training data while the protoype projection uses unaugmented training samples and the `SPARROW` evaluation protocol is applied on all unaugmented training and test samples.

Data Preprocessing Preprocessing of the samples from the CUB dataset follows the approach from Chen *et al.* [1] and consists of the following steps in the given order:

- The birds are cropped from the background by bounding boxes which are provided by the dataset.
- All samples are rescaled to size $S_{\text{raw}} = (224, 224)$.
- Data augmentation is applied on all training samples which consists of rotating, skewing, shearing, randomly distorting and flipping the images.
- All samples are normalized by the mean and standard deviation of the training set.

[1] Parts are: back, beak, belly, breast, crown, forehead, left eye, left leg, left wing, nape, right eye, right leg, right wing, tail, throat. Not all parts are visible in every sample.

Context	Parameter	Value	Explanation
Data (CUB)	n_{train}	5994	
	n_{test}	5794	
	N	11788	
	n_{aug}	215784	
	$\rho_{\text{train-val}}$	0.9	
	n	194206	
	n_{val}	21578	
	C	200	
	T	15	
	S_{raw}	$(224, 224)$	
Model	S_{pt}	$(1, 1)$	
	m	$\{1000, 2000, 4000\}$	Number of prototypes used in the experiments
Training	$\lambda_{\mathcal{L}^{\text{CrsEnt}}}$	1	
	$\lambda_{\mathcal{L}^{\text{Clst}}}$	1	
	$\lambda_{\mathcal{L}^{\text{Sep}}}$	0	
	$\lambda_{\mathcal{L}^{\text{AS}}}$	0 / 1	for baseline / own method
	$\lambda_{\mathcal{L}^{\text{PSD}}}$	0 / 100	for baseline / own method
	$\lambda_{\mathcal{L}^{L1}}$	1e−2	
	batch size	32	Number of samples per mini-batch in a training iteration
	optimizer	Adam [4]	
	lr_{base}	1e−4	Learning rate in the base network (cf. table 2)
	$\text{lr}_{\text{add-on}}$	1e−4	Learning rate in the add-on network (cf. table 2)
	$\text{lr}_{\text{prototype}}$	3e−3	Learning rate in the prototype layer (cf. table 2)
	$\text{decay}_{\text{base}}$	1e−3	Weight decay of the Adam opt. in the base network
	$\text{decay}_{\text{add-on}}$	1e−3	Weight decay of the Adam otp. in the add-on network
	pat_{warm}	1	Warm start early stopping patience in epochs
	pat_{main}	3	Main training early stopping patience in epochs
	$\text{delta}_{\text{warm}}$	1e−2	Warm start early stopping criterion
	$\text{delta}_{\text{main}}$	5e−3	Main training early stopping criterion
	$\text{epochs}_{\text{ll}}$	3	Epochs of convex optimization of the last layer

Table 3: **Choices for Hyperparameters**. The meaning of the symbols is explained in table 1.

With $\mathcal{L}^{\text{AS}}$	With $\mathcal{L}^{\text{PSD}}$	$\text{epochs}_{\text{warm}}$	$\text{epochs}_{\text{main}}$	time [h]	acc_{val} [%]
False	False	10	6	6.88	98.6
False	True	9	9	8.00	99.0
True	False	10	19*	72.23*	90.7*
True	True	10	8	34.42	97.2

Table 4: **Convergence and training duration** in epochs and hours with the hardware as described in section 5 and hyperparameters in table 3. The time calculation includes all training steps (prototype learning, prototype push, last layer optimization), phases (warm start and main training) and the performance evaluation. acc_{val} is the final validation accuracy. Entries marked with * are not final since the respective experiment did not finish until the submission of this supplementary material.

Hyperparameter Optimization An overview over all hyperparemeters is shown in table 3. Most parameters are oriented on the work of Chen *et al.* [1]. In order to tune the static weights of the loss components $\lambda_{\mathcal{L}^\nu}$ ($\nu \in \{\text{CrsEnt}, \text{Clst}, \text{Sep}, \text{AS}, \text{PSD}\}$), we performed two rounds of subsequent random search and Bayesian Optimization (BO) runs. For BO we used an Upper Confidence Bound acquisition function and a Gaussian Process with a Matern kernel as a surrogate. We optimized for a measure of balance between performance and interpretability for which we used

$$\text{opt} = \frac{2 \cdot \text{acc}_{\text{val}} + \text{ptd} + \text{ptf}}{4}, \tag{3}$$

with the validation accuracy acc_{val} and the *ptd* and *ptf* measures as defined in section 2 in the main paper. We reduced the time demand of the tuning process by randomly reducing the number of classes to 20 in each experiment and ran each experiment until early stopping.

In both optimization rounds, we ran 30 initial random searches which seeded 45 iterations of BO. In the first round, we selected values for each loss component weight from an even distribution of 100 values on a logarithmic scale over the interval $[0.01, 100]$. We found that for the top performing runs, $\lambda_{\mathcal{L}^{\text{AS}}}$ was typically smaller than the weight of the other loss components except for $\lambda_{\mathcal{L}^{\text{Sep}}}$ which was typically optimized to $\lambda_{\mathcal{L}^{\text{Sep}}} = 0.01$ with all other loss component weights being at least two orders of magnitude larger. We performed a second optimization round for fine-tuning in which we set $\lambda_{\mathcal{L}^{\text{Sep}}} = 0.01$ and $\lambda_{\mathcal{L}^{\text{AS}}} = 1$ and selected the other loss component weights from an even distribution of 15 values on a logarithmic scale over the interval $[1, 100]$. Amongst the top performing runs we opted for a simple choice for the set of remaining parameters ($\lambda_{\mathcal{L}^{\text{CrsEnt}}} = 1$, $\lambda_{\mathcal{L}^{\text{Clst}}} = 1$, $\lambda_{\mathcal{L}^{\text{PSD}}} = 100$). We did a final round of experiments where we investigated if we could do without one of the loss components and actually found that setting $\lambda_{\mathcal{L}^{\text{Sep}}} = 0$ slightly improved our optimization goal (cf. eq. 3). This might be because the newly introduced *PSD* loss which tries to keep prototypes close to samples of the same class fulfills a similar purpose (i.e. keeping samples away from prototypes of other classes).

Training hardware Experiments were performed on single NVIDIA GeForce RTX 1080 Ti GPUs with 11 GB GDDR6 of graphics memory.

Convergence and training duration We analyzed how the training duration varies based on inclusion or exclusion of the new loss components $\mathcal{L}^{\text{AS}}$ and $\mathcal{L}^{\text{PSD}}$. The results are reported

Prototypes	Total runtime [h]
1000	5.25
2000	10.57
4000	19.35

Table 5: **Time demand for the SPARROW evaluation protocol** for different numbers of prototypes with the hardware as described in section 5 and hyperparameters in table 3.

in table 4. The ProtoPNet model without any additional loss components (first line) has the shortest runtime until convergence. Including the $\mathcal{L}^{\text{PSD}}$ loss component (second line) only slightly increases time to convergence. However, including the $\mathcal{L}^{\text{AS}}$ loss component (third line) but not the $\mathcal{L}^{\text{PSD}}$ loss component leads to a very slow convergence and as a result to a greatly increased time demand until early stopping. This experiment did not finish until the submission of this supplementary material but the slow convergence is clearly visible. This slow convergence is mitigated to a reasonable extent by also including the $\mathcal{L}^{\text{PSD}}$ loss component (last line) which is our proposed method. Its runtime is about five times as high as that of the original ProtoPNet model (first line).

Optimizing the time-demand of our method and developing new methods with reduced time-demand but comparable explanatory capacity (measured by the SPARROW evaluation protocol) are goals we want to pursue in future work.

Time demand for the SPARROW evaluation protocol The time-demand of the SPARROW evaluation protocol for different numbers of prototypes is shown in table 5. It can be seen that it increases about linearly with the number of prototypes. These numbers should however be seen as an easy to reach upper limit since we did not yet optimize our implementation for performance. We see this as a goal for future work.

References

[1] Chaofan Chen, Oscar Li, Daniel Tao, Alina Barnett, Cynthia Rudin, and Jonathan K Su. This looks like that: deep learning for interpretable image recognition. In *Advances in neural information processing systems (NeurIPS)*, 2019.

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015.

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[4] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[5] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style,

high-performance deep learning library. In *Advances in Neural Information Processing Systems (NeurIPS) 32*. 2019.

[6] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The Caltech-UCSD Birds-200-2011 Dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011.

C ALPEC: A Comprehensive Evaluation Framework and Dataset for Machine Learning-Based Arousal Detection in Clinical Practice

The full text of the following publication is included in this appendix:

Stefan Kraft, Andreas Theissler, Vera Wienhausen-Wilke, Philipp Walter, Gjergji Kasneci, and Hendrik Lensch. ALPEC: A comprehensive evaluation framework and dataset for machine learning-based arousal detection in clinical practice. In Xuhai Orson Xu, Edward Choi, Pankhuri Singhal, Walter Gerych, Shengpu Tang, Monica Agrawal, Adarsh Subbaswamy, Elena Sizikova, Jessilyn Dunn, Roxana Daneshjou, Tasmie Sarker, Matthew McDermott, and Irene Chen, editors, *Conference on Health, Inference, and Learning, UC Berkeley, Berkeley, USA, 25-27 June 2025*, volume 287 of *Proceedings of Machine Learning Research*, pages 395–429. PMLR, 2025. URL https://proceedings.mlr.press/v287/kraft25a.html

ALPEC: A Comprehensive Evaluation Framework and Dataset for Machine Learning-Based Arousal Detection in Clinical Practice

Stefan Kraft STEFAN.KRAFT@IT-DESIGNERS.DE
IT-Designers Gruppe & University of Tübingen, Germany

Andreas Theissler HTTPS://ORCID.ORG/0000-0003-0746-0424
Aalen University of Applied Sciences, Germany

Vera Wienhausen-Wilke V.WIENHAUSEN-WILKE@KLINIKUM-ESSLINGEN.DE
Klinikum Esslingen, Germany

Philipp Walter PHILIPP.WALTER@IT-DESIGNERS.DE
IT-Designers Gruppe, Germany

Gjergji Kasneci GJERGJI.KASNECI@TUM.DE
Technical University of Munich, Germany

Hendrik Lensch HENDRIK.LENSCH@UNI-TUEBINGEN.DE
University of Tübingen, Germany

Abstract

Detecting arousals during sleep is crucial for diagnosing sleep disorders, yet the adoption of Machine Learning (ML) in clinical practice is hindered by a mismatch between clinical protocols and ML methods. Clinicians typically annotate only arousal onsets, whereas ML approaches conventionally rely on annotations for both the beginning and end. Moreover, no standardized evaluation methodology exists that is tailored to the specific needs of arousal detection in clinical practice. We address these challenges by proposing a novel post-processing and evaluation framework – Approximate Localization and Precise Event Count (ALPEC) – which optimizes arousal detectors to reflect operational priorities. We further advocate focusing on arousal onset detection and assess the impact of this on current training and evaluation schemes, addressing associated simplifications and challenges. Finally, we introduce the novel Comprehensive Polysomnographic (CPS) dataset that reflects the aforementioned clinical annotation constraints and includes modalities absent from existing datasets, demonstrating the benefits of leveraging multimodal data for arousal onset detection. Our contributions significantly advance the integration of ML-based arousal detection into clinical settings, narrowing the gap between technological advancements and clinical requirements.

Data and Code Availability This paper introduces the CPS dataset (Kraft et al., 2024; Goldberger et al., 2000) which we collected during clinical practice from 2021-2022 monocentrically at Klinikum Esslingen, Germany. It is released on the PhysioNet platform and is accessible under the PhysioNet Credentialed Health Data License 1.5.0. We also utilize the 2018 PhysioNet Challenge Dataset (Ghassemi et al., 2018; Goldberger et al., 2000) which is also available on the PhysioNet repository.

While the code for our training and evaluation procedures is proprietary, besides formalizing ALPEC (Section 3.2), we provide a schematic scheme comparison (Appendix E), pseudo-code (Appendix F), and instructions for running baseline training and evaluation schemes (Sections 3.1 and 3.3). The code for these baselines is publicly available[1]. The official documentation of the CPS dataset on PhysioNet as well as the supplementary material include Croissant (Akhtar et al., 2024) specifications, code, and instructions for data loading.

Institutional Review Board (IRB) Our study protocol was approved by the ethics committee of the Landesärztekammer Baden-Württemberg, Germany, on 2020-10-21 (committee number F-2020-105).

1. DeepSleep 2.0: https://github.com/rfonod/deepsleep2 (MIT license); sktime: https://doi.org/10.5281/zenodo.3749000 (BSD-3-Clause license)

1. Introduction

Arousals are short-term biological activation processes during sleep and wakefulness that elevate the organism from a lower to a higher state of mental and physical activity (Raschke and Fischer, 1997). Frequent arousals during sleep disrupt deep sleep stages and REM sleep, compromising the restorative function of sleep and causing fragmentation (Wetter et al., 2012). Arousals are indicative of several sleep disorders, with obstructive sleep apnea (OSA) being the most prevalent breathing-related sleep disorder. OSA, characterized by partial or complete obstructions of the upper airway, results in oxygen desaturation and frequent arousals (Wetter et al., 2012). This disorder is a significant public health concern, with prevalence estimates around 20% in men and 17% in women, and is linked to severe health outcomes like hypertension, cardiovascular disease, and increased mortality risk (Franklin and Lindberg, 2015; Wetter et al., 2012; Punjabi, 2008).

Detecting arousals is a routine task in polysomnographic (PSG) examinations, which involve comprehensive recording and analysis of various physiological parameters such as brain waves, blood oxygen levels, breathing, eye and leg movements during sleep, conducted in sleep laboratories. However, the diversity of equipment, software, and protocols across laboratories poses significant challenges for developing universally applicable Machine Learning (ML) models for arousal detection (Anido-Alonso and Alvarez-Estevez, 2023). Even among laboratories using the same equipment, differences in settings and protocols complicate the development of general-purpose ML-based detectors. The lack of large-scale time series datasets, absence of clear evaluation metrics, and limited consensus on theoretical and practical understanding of time series further impede progress (Garza and Mergenthaler-Canseco, 2023). A significant drop in performance of current sleep stage classification models on data from different sleep laboratories highlights the need for arousal detection models trained on data specific to the clinical environments where they will be used (Anido-Alonso and Alvarez-Estevez, 2023).

According to the American Association of Sleep Medicine (AASM), arousals are defined as abrupt changes in EEG frequency lasting at least 3 seconds following 10 seconds of sleep (Berry et al., 2012). Although research suggests that the duration of arousals may have clinical significance (Schwartz and Moxley, 2006; Shahrbabaki et al., 2021), this finding has not been integrated into clinical practice. This is evident from the AASM guidelines, which recommend reporting solely the number of arousals and the arousal index, a measure that quantifies the frequency of arousals per hour of sleep, for PSG diagnostics (Berry et al., 2012). This aligns with our CPS dataset, where nearly all annotated arousals are three seconds in duration, which is the default setting of the clinical annotation software. This limitation implies that only the onset annotations are practically useful. Current training approaches for arousal detection, which rely on full annotations encompassing both the start and end of each event, diverge from this clinical practice. Consequently, they are misaligned with the clinical reality, hindering their application in real-world settings.

Our **first main contribution** is advocating for a shift in focus from full event detection to detecting arousal onsets to better align with clinical needs. We explore the implications on various training methodologies for binary event detection, addressing both simplifications and emerging challenges. Conversely, aligning clinical practices with ML model requirements without evident patient benefits would unnecessarily burden sleep laboratories, increasing operational challenges such as the scarcity of trained scorers and long patient wait times.

In addition to task misalignment, the fragmented landscape of evaluation methodologies lacks the operational utility required for real-world healthcare settings. Our **second main contribution** is aligning the performance evaluation of arousal detection systems for decision support rather than autonomous decision-making for which we advocate using the F2 score. A clinical decision support system (CDSS) augments the clinical workflow by highlighting potential arousal occurrences, enabling clinicians to focus on the most relevant segments of sleep recordings. Clinicians still retain final responsibility for diagnostic conclusions, but the system streamlines their work by pinpointing events needing review. This approach adheres to ethical standards in healthcare (Fawzy et al., 2023) and anticipates evolving regulatory requirements, such as the forthcoming EU AI Act, which mandates human oversight for AI systems in critical areas like healthcare (Madiega, 2021).

This aligned evaluation approach is integral to ALPEC – a post-processing and performance evaluation framework that constitutes our **third main contribution**. ALPEC is embedded within a rel-

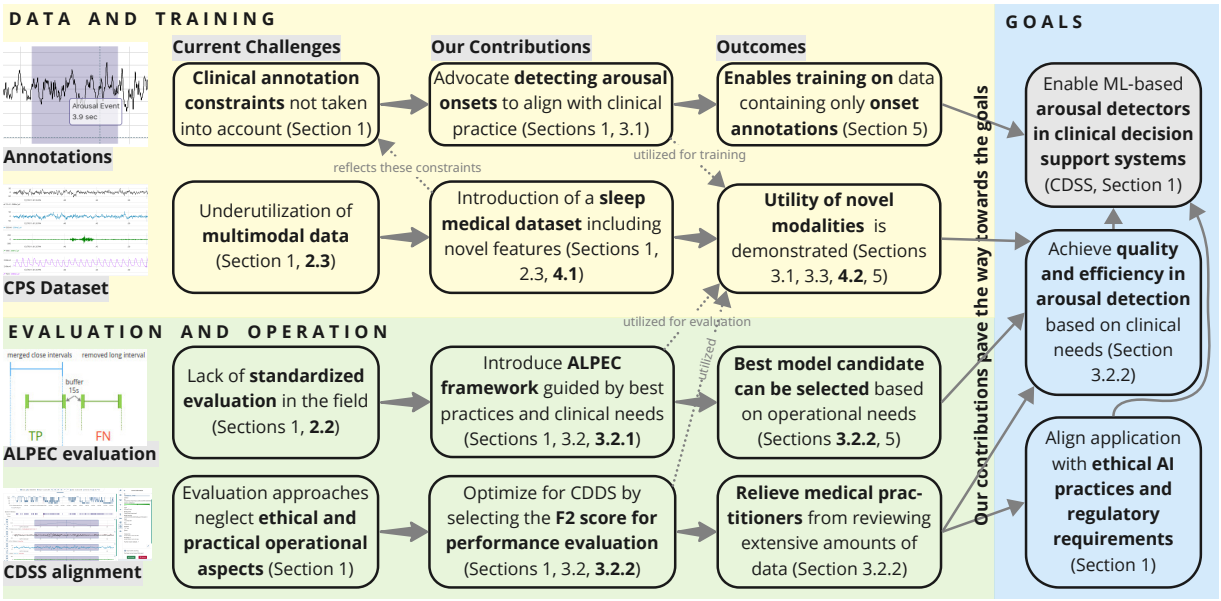


Figure 1: **Contextualization of our contributions** within the broader landscape of challenges in the field, highlighting the outcomes and underlying goals of our work. Key sections for each topic are referenced.

evant taxonomy, overcomes key conceptual limitations, and adheres to best practices. It is the first framework designed to standardize performance evaluation in the field of arousal detection, and it is tunable to the needs of other applications and domains that require precise event count detection. We provide a thorough comparison of our framework with existing evaluation methodologies.

Lastly, as our **fourth main contribution**, we are excited to introduce the Comprehensive Polysomnography (CPS) dataset, a curated, feature-rich collection unique for its extensive channels and novel beat-by-beat blood pressure annotations. This dataset aims to advance ML models in sleep disorder diagnostics. We present the first study on this dataset, demonstrating enhanced arousal detection capabilities through multimodal data, advancing the integration of previously underutilized data modalities.

Figure 1 provides a structured contextualization of our contributions and may serve as a guide for navigating the paper.

2. Related work

In this section, we embed our work into related work on arousal detection, highlighting training approaches (Section 2.1), diverse evaluation practices (Section 2.2), and notable datasets (Section 2.3).

2.1. Methods for arousal detection

Table 1 provides an overview of various approaches for arousal detection, highlighting its close association with sleep stage classification, where the primary goal is to determine the sleep stage for each 30-second epoch of a polysomnographic recording. Similar to sleep stage classification, for arousal detection, the data is typically segmented into N consecutive windows of fixed length s, with s either optimized as a hyperparameter or fixed at 30 seconds (Li et al., 2018; Phan et al., 2019) or other durations (Kuo et al., 2023). Overlapping windows are frequently employed to better capture arousal events, sometimes extended for evaluation as well (Badiei et al., 2023; Li et al., 2018). The definition of when a window signifies an arousal event often introduces another layer of complexity, sometimes determined by majority voting within the window or by the presence of at least one arousal label (Kuo et al., 2023), which may lead to another hyperparameter (Li et al., 2018).

We perform window-based segmentation only for baseline comparisons, as our primary focus is on continuous segmentation. For this, we build on the

Table 1: **Comparison of Related Work**. This comparison showcases the diversity of methodologies utilized in the field. The datasets are explained in Section 2.3. For counting *#Modalities*, multiple EEG channels are considered a single modality, while derived channels or features are counted separately. Notations: *Seg.* denotes segmentation methods; *(AU)PRC* and *(AU)ROC* indicate that either the curve, the area under the curve, or both are reported.

Authors	Task		Seg.		Dataset					Evaluation measures									Modalities					
	Arousal	Sleep Stage	Windowed	Pointwise	2018 Phys	SHHS	MESA	CPS (our)	Unpublished	(AU)PRC	(AU)ROC	Accuracy	Sensitivity	Specificity	Precision	Recall	β for F_β	Cohen's κ	EEG	EMG	ECG	EOG	Other	#Modalities
Badiei et al. (2023)	**x**	**x**	x		x	x					x	x	x						x		x			2
Foroughi et al. (2023)	x		x		x						x	x	x	x					x					1
Li et al. (2018)	x		x		x					**x**	**x**								x	x				2
Kuo et al. (2023)	x		x						x	x	x	x			x	x	1						x	18
Miller et al. (2018)	x			**x**	x					**x**	**x**								x	x	x		x	7
Howe-Patterson et al. (2018)	**x**	**x**		**x**	x					**x**	**x**								x	x		x	x	7
Li and Guan (2021)	**x**	**x**		**x**	x	x				**x**	**x**								x	x	x	x	x	8
Fonod (2022)	x			**x**	x					**x**	**x**								x	x	x	x	x	8
Zan and Yildiz (2023)	**x**	**x**		**x**		x	x			x	x	x			x	x	1	x	x					1
Ehrlich et al. (2024)	x			**x**		x			x	x							1		x	x		x		3
our	x		x	**x**	x			**x**							x	x	**2**		x	x	x	x	x	**39**

methodological foundation of the *DeepSleep* architecture, which employs a Fully Convolutional Neural Network (FCN) with a U-Net architecture to process extensive polysomnographic signals continuously (Li and Guan, 2021). This model differs from windowing approaches as it handles the entire dataset as a single sequence, where each point is evaluated within the context of its receptive field. This comprehensive approach to arousal detection offers several advantages over traditional window-based methods: It eliminates the need for multiple hyperparameters, does not require manual feature extraction, supports an end-to-end process, processes multimodal data natively, and leverages extensive temporal contexts to capture interactions across various timescales (Li and Guan, 2021). This has spurred a growing body of research pursuing similar comprehensive methodologies, as documented in Table 1.

2.2. Current state of evaluating arousal detection models

Evaluating arousal detection models is challenging due to the diversity of methodologies (e.g., pointwise vs. window-based evaluations) and the absence of standardized evaluation protocols (Foroughi et al., 2023; Badiei et al., 2023). This diversity – reflected in Table 1 – is compounded by the wide range of performance metrics employed. In practice, window-based evaluations dominate, as training typically favors window-based classification (WBC) over continuous segmentation (CS). For example, Zan and Yildiz (2023) use CS for both sleep stage classification and arousal detection in a multitask setup, yet still evaluate by applying 30-second non-overlapping sliding windows with labels assigned via majority vote or the mere presence of an arousal indicator. We employ pointwise and window-based evaluation methods only as baselines. In contrast, we propose an event-based evaluation approach that treats each contiguous segment of predictions as a single event to be compared with the ground-truth. Our method improves the segment-wise f-score introduced by Hundman et al. (2018) for time series anomaly detection (TSAD) and is also applicable when only event onsets are annotated. Notably, Ehrlich et al. (2024) also employ event-based evaluation for arousal detection with similar strategies – such as label adaptations, event merging, temporal tolerance, and a

specific counting scheme. However, our contribution goes further: Our approach (ALPEC) is embedded within a relevant taxonomy of evaluation metrics that was proposed by Sørbø and Ruocco (2023) for TSAD, which advocates tailoring the selection of evaluation metrics to operational needs. ALPEC addresses the requirements of clinical decision support systems and may serve to standardise the evaluation of arousal detectors, ending methodological fragmentation and ensuring comparability of results across studies.

2.3. Datasets and data modalities for arousal detection

Prominent datasets in sleep research include the 2018 PhysioNet Challenge dataset (Ghassemi et al., 2018; Goldberger et al., 2000), the Sleep Heart Health Study (SHHS) (Quan et al., 1997; Zhang et al., 2018), and the Multi-Ethnic Study of Atherosclerosis (MESA) (Chen et al., 2015; Zhang et al., 2018). Another notable and more recent collection of pediatric sleep data is the NCH Sleep DataBank (Lee et al., 2021, 2022; Goldberger et al., 2000). Like our CPS dataset, these datasets offer extensive PSG data and (in case of SHHS and MESA) patient information collected through standardized questionnaires such as sleep and restless legs questionnaires, the Pittsburgh Sleep Quality Index (PSQI), and the Epworth Sleepiness Scale (ESS). As indicated in the last column of Table 1, many studies utilize only few modalities. Those that employ more, such as Kuo et al. (2023) on an unpublished dataset, often perform feature engineering to derive additional features. Others, like Li and Guan (2021) and Fonod (2022), are constrained by the number of available channels in the 2018 PhysioNet Challenge dataset.

Our CPS dataset, in contrast, offers 17 raw channel modalities and numerous derived features by use of the DOMINO expert software from SOMNOmedics GmbH, featuring innovative modalities such as pulse transit time and beat-by-beat blood pressure estimations. The potential of these modalities in sleep diagnostics is supported by various studies (Misaka et al., 2020; Argod et al., 1998; Pitson et al., 1998, 1994), and we aim to further investigate their impact on arousal detection in an ongoing clinical study (Wienhausen-Wilke and Kraft, 2024). Additional highlights of the CPS dataset include annotations indicating whether arousals were first detected in the EEG or as a consequence of other physiological changes, along with detailed medical outcomes such as sleep diagnoses, Baveno classification, and T90 value. Further details on the CPS dataset are provided in Appendix H.

3. Methodologies

We present our core methodologies for event detection and performance evaluation, grounded in real-world considerations to enable robust clinical decision support for arousal detection in sleep medicine. We first describe our main approach for arousal onset detection in Section 3.1, present our novel framework ALPEC in Section 3.2, and then introduce multiple baseline approaches in Section 3.3.

3.1. Arousal detection by continuous segmentation

We start by detailing our main approach for arousal onset detection. We adopt the DeepSleep architecture, which facilitates continuous segmentation of data into distinct classes (Li and Guan, 2021). We build on an optimized version of this architecture, proposed by Fonod (2022), under MIT license, which reduces the U-Net's depth from 11 to 5 layers, substantially decreasing computational demands while maintaining comparable performance. This streamlined model processes all data points from multi-channel sleep recordings simultaneously, eliminating the need for window-based classification. It translates these inputs into sleep arousal scores for each data point using a binary cross-entropy (BCE) loss function. We refine this approach by employing a weighted BCE loss, adjusting the loss contribution of each data point by inversely weighting it according to the frequency of the arousal class within the subject's data, addressing class imbalance.

Since detecting singular arousal onset points does not work well with the DeepSleep approach (see Section 4.1), we modify the ground-truth annotations to mark intervals of length l around each arousal onset as positive. We select $l = 10$ seconds, aligning with arousal scoring rules that require at least 10 seconds of stable sleep between distinct arousal events, ensuring the created ground-truth intervals do not overlap (Berry et al., 2012). We call this approach *interval-based onset detection*. During inference, the DeepSleep model outputs probability scores $p_i(\mathbf{x})$ for each data point i. To smooth these outputs for reducing false detections, we apply an averaging filter over a smoothing window of $w = 3$ seconds per point.

3.2. ALPEC: Approximate localization and precise event count framework for post-processing and performance evaluation

Sørbø and Ruocco (2023) rightly state that there is no universally correct set of metrics for any specific task; however, using inappropriate metrics can lead to suboptimal decisions when selecting algorithms for productional use.

Guided by their taxonomy, we developed the Approximate Localization and Precise Event Count (ALPEC) framework to address the need for standardized performance evaluation in arousal detection that align with the operational goals of clinical practice. This is crucial for making informed decisions about the deployment of arousal detection algorithms in real-world clinical settings.

We first formally introduce the ALPEC procedure in Section 3.2.1, before discussing its rationale and our hyperparameter choice in Section 3.2.2. A schematic embedding of ALPEC into various training and evaluation schemes is provided in Appendix E (Figure 2), an algorithmic description in Appendix F, and a table of notation in Appendix A (Table 9).

3.2.1. Formal description of ALPEC

We now formally introduce the procedure of our post-processing and performance evaluation framework ALPEC.

ALPEC is compatible with both window-based classification (WBC) and continuous segmentation (CS) approaches to arousal detection. When using CS (see Section 3.1), we start with a probability score $p_{i,\nu}(\mathbf{x})$ for each data point $i = 1, ..., n$ from the measurement data $\mathbf{x}$ and each subject ν with $\nu = 1, ..., |D|$ in the dataset D, where $n = 2^{23}$ is the padded fixed number of data points for each input channel. Alternatively, if we use WBC (see Section 3.3), we start with probability scores $p_{\eta,\nu}(\mathbf{x})$ or binary class predictions $c_{\eta,\nu}(\mathbf{x})$, where $\eta = 1, ..., N$, and the data is divided into N windows of equal length s.

When we have probability scores, i.e., $p_{\xi,\nu}(\mathbf{x})$ with $\xi \in \{i, \eta\}$, we apply a threshold t_k to the scores to obtain binary class predictions $c_{\xi,\nu,k}(\mathbf{x})$, where thresholds t_k are selected from 0 to 1 in steps of 0.01, i.e., $k = 0, 1, ..., 100$:

$$c_{\xi,\nu,k}(\mathbf{x}) = \begin{cases} 1 & \text{if } p_{i,\nu}(\mathbf{x}) \geq t_k \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

In the case of WBC, we resample the window-based predictions to pointwise predictions $c_{i,\nu}(\mathbf{x})$ by assigning the prediction of the window to all data points within the window:

$$c_{i,\nu}(\mathbf{x}) = c_{\eta,\nu}(\mathbf{x}) \text{ for } i \in [(\eta - 1) \cdot s + 1, \eta \cdot s] \tag{2}$$

At this point, both starting points (continuous segmentation and window-based classification) are synchronous.

Interval Merging Next, we merge predictions less than δ seconds apart. For ease of notation, we temporarily drop the indices k and ν and parameter $\mathbf{x}$. At first, for the sequence of binary target values $C = (c_1, c_2, ..., c_n)$, we identify the start and end indices of each predicted interval P in C as P^{start} and P^{end}, respectively, where an interval starts at index i if $c_i = 1$ and $c_{i-1} = 0$ and ends at index j if $c_j = 1$ and $c_{j+1} = 0$ if all intermediate predictions $c_{i+1}, ..., c_{j-1}$ are equal to 1. For *arousal onset detection*, the distance is calculated based on the maxima of the scores of two consecutive predicted intervals. For each identified interval P, we find the index m within the interval that maximizes the score p_m:

$$m = \arg\max_{m \in [P^{\text{start}}, P^{\text{end}}]} s_m \tag{3}$$

Two intervals P_1 and P_2, with maximum score indices m_1 and m_2, are merged if $|m_1 - m_2| < \delta \cdot f$, where f is the sampling frequency of the data. In the case of *full event detection*, we merge two intervals based on their start and end points, i.e., if $|P_1^{\text{start}} - P_2^{\text{end}}| < \delta \cdot f$. The merged interval P^{merged} then extends from P_1^{start} to P_2^{end}. For the merged sequence C^{merged} we set:

$$c_i^{\text{merged}} = \begin{cases} 1 & \text{if } i \in \text{any} P^{\text{merged}} \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

Matching Predictions and Ground-Truth The final step is to compare the predicted intervals P with the ground-truth intervals G with start end points G^{start} and G^{end}. In the case of *point-based onset detection*, $G^{\text{start}} = G^{\text{end}}$, meaning the ground-truth intervals are points in time. Our ALPEC framework, however, remains generic and can be used for any length of ground-truth intervals, thus supporting both *interval-based onset detection* and *full event detection*.

Next, we once again tailor the evaluation method to the specificities of the task by introducing two two

key **approximate localization** components: First, we define a **maximum duration** d for the predicted intervals, which we will utilize shortly. Second, we extend all ground-truth intervals with a **temporal tolerance buffer** b^{before} on the left and b^{after} on the right side of the interval:

$$G^{\text{start,ext}} = \max(0, G^{\text{start}} - b^{\text{before}} \cdot f) \tag{5}$$

$$G^{\text{end,ext}} = \min(n, G^{\text{end}} + b^{\text{after}} \cdot f) \tag{6}$$

Now, we can compare predicted and ground-truth intervals to calculate true positive (TP), false positive (FP), and false negative (FN) counts, meeting the requirement of **precise event counts**. A TP is counted when a predicted interval P overlaps with a ground-truth interval G^{ext} extended by the buffer, so that $P^{\text{start}} \leq G^{\text{end,ext}}$ and $P^{\text{end}} \geq G^{\text{start,ext}}$ and $P^{\text{end}} - P^{\text{start}} \leq d \cdot f$, restricting the maximum duration of a predicted interval. A FP is counted if a P does not overlap with any G^{ext}. A FN is counted if a G^{ext} does not overlap with any P. If multiple P overlap with a single G^{ext}, we count one TP for the first overlap and each additional overlap as a FP. If a single P overlaps with multiple G^{ext}, we count one TP for the first match and each unmatched G^{ext} as a FN.

Performance Evaluation Next, we use established formulas to calculate the $\text{Precision}_{\nu,k}$, $\text{Recall}_{\nu,k}$, and $\text{F2}_{\nu,k}$ score for each subject ν and each threshold t_k from the TP, FP, and FN counts, reintroducing the indices.

We use the F2 score for selecting the optimal threshold. The first step is to calculate the micro average $F2_k^{\text{train}}$ score across subjects $v^{\text{train}} = 1, ..., |T|$ of the training set T for each threshold t_k. as $\text{F2}_k^{\text{train}} = \frac{1}{|T|} \sum_{\nu \in v^{\text{train}}} \text{F2}_{\nu,k}$. From this, we determine the optimal threshold t_k^{opt} with $k^{\text{opt}} = \arg\max_k \text{F2}_k^{\text{train}}$ which maximizes the average $F2$ score on the training set. We then obtain the mean Precision, Recall, and F2 score across subjects for the optimal threshold t_k^{opt}. We select the mean F2 score as the final metric for performance evaluation and report the mean Precision and Recall as auxiliary metrics for additional insights (performance analysis).

3.2.2. Rationale and Hyperparameter Choice for Arousal Detection

ALPEC shares several similarities with existing metrics. First, it is most similar to the segment-wise f-score (Hundman et al., 2018), as both approaches focus on evaluating segment overlaps rather than pointwise predictions. Second, like some existing methods, ALPEC employs a **temporal tolerance buffer** around ground-truth intervals (Scharwächter and Müller, 2020), that we set to $b^{\text{before}} = b^{\text{after}} = 15s$, corresponding to the typical length of one 30-second epoch as viewed by medical scorers. The use of this buffer addresses potential temporal inaccuracies in arousal event annotations and is backed by the irrelevance of precise annotations in current clinical practice, leading to the **approximate localizaton** requirement, making the evaluation process more robust and clinically relevant. Third, the integration of ALPEC into the taxonomy by Sørbø and Ruocco (2023) demonstrates its alignment with established categories in time series anomaly detection. ALPEC metrics classify as *binary*, since thresholding does not manipulate prediction scores, and *redefined counting-based*, as they involve comparing intervals rather than evaluating pointwise. Additionally, they exhibit intrinsic insensitivity to true negatives and a valuation property of time tolerance.

ALPEC also introduces several significant differences: First, ALPEC **merges close predicted intervals**, where we set $\delta = 10s$, in line with clinical guidelines requiring at least 10 seconds of stable sleep between arousals (Berry et al., 2012). Second, ALPEC imposes a **maximum duration** on predicted intervals, which we set to $d = 60s$, ensuring they do not exceed practical lengths. This restriction intends to prevent ambiguities during human review and maintain the clinical relevance of arousal timing relative to sleep stages (that are scored in 30-second epochs) which is important for physicians in sleep medicine. Third, the method of **precise event counting** ensures that only one TP is counted per predicted interval, which is essential for clinical utility. If a predicted interval spans multiple ground-truth intervals, it results in multiple FNs unless each ground-truth interval is uniquely matched to a predicted interval. This approach avoids the pitfall of inaccurately rewarding temporal extension of predicted intervals, a limitation of segment-wise f-score methods (Sørbø and Ruocco, 2023).

Finally, unlike most current approaches to arousal detection, which typically rely on the F1 score or limit their reports to AUPRC or AUROC without a clear consensus on the most appropriate metric (see Table 1), ALPEC follows Sørbø and Ruocco (2023) in advocating for a context-aware selection of metrics

tailored to the operational environment and model selection process. Relying solely on AUPRC or AUROC is inadequate for **clinical decision-making**, as these metrics do not fully capture how model predictions translate into actionable outcomes. For applications, where reliability is crucial, task-specific threshold analysis is preferable, as it directly reflects operational trade-offs. Also, recent research highlights potential biases for AUPRC and questions its applicability in high-stakes decision-making (McDermott et al., 2024). Specifically, AUPRC tends to show high variance in imbalanced datasets with few positive samples, such as arousal events in our case, making it an unreliable criterion for model selection. Therefore, we acknowledge the utility of these metrics in performance analysis but do not consider them sufficient for performance evaluation and model selection. AI-based clinical decision support systems (CDSS) in healthcare aim not only to improve the quality of outcomes but also to enhance the efficiency of medical practitioners' work (Magrabi et al., 2019; Vasey et al., 2022). Given the need to relieve medical practitioners from reviewing extensive amounts of data – in our case, night-long recordings – it is crucial for CDSS to highlight the most pertinent data sections. This objective is best met by ensuring that AI predictions minimize missed arousals (false negatives), allowing human reviewers to efficiently address any false positives. The F2 score is particularly well-suited for this purpose, as it explicitly prioritizes recall over precision, reducing the likelihood of missed arousal events.

3.3. Baseline approaches

In this work, we also explore traditional window-based classification (WBC) methods to provide a comparative baseline for arousal detection. These WBC approaches are evaluated using both standard window-based evaluation and our ALPEC framework. This helps in evaluating the effectiveness of our proposed approach against established techniques.

We employ several classical univariate models from the sktime library (Löning et al., 2019) (BSD 3-Clause License), utilizing them with their standard configurations. For each arousal onset in the training set, we construct a 30-second window centered randomly around the onset point to enhance generalizability. This random alignment aims to mimic the variable alignment of arousal onset points during inference across non-overlapping windows covering the entire series of a subject. To address the challenge of class imbalance, we select an equal number of negative-class windows randomly for each subject during training. The evaluation then proceeds with standard WBC, dividing the test set data into consecutive non-overlapping windows of the same 30-second length used in training. Adopting window-based onset detection simplifies the windowing approach by reducing the need for overlapping windows or complex voting schemes. Instead, we apply the *presence* criterion to determine the class of the windows, aiming to approximate the original class distribution in the test set. Additionally, we ensure robustness by conducting cross-subject validation, where training and testing sets include distinct subjects. Furthermore, we conduct baseline experiments where CS approaches are evaluated using both traditional window-based evaluation and ALPEC for comparative analysis. This involves creating windows for ground-truth and prediction as described above.

4. Results

We perform an experimental comparison of training and evaluation schemes based on the 2018 PhysioNet Challenge Dataset in Section 4.1 and our CPS dataset in Section 4.2. For an illustration of the schemes, refer to Appendix E (Figure 2). Ablation studies on hyperparameter choices are reported in Appendix A.

4.1. Experiments on the 2018 PhysioNet Challenge dataset

The 2018 PhysioNet Challenge Dataset, licensed under the Open Data Commons Attribution License v1.0, is a notable publicly available resource that includes polysomnographic (PSG) data from 1,983 patients at Massachusetts General Hospital's Sleep Lab, with labels provided for 994 subjects (Ghassemi et al., 2018; Goldberger et al., 2000). It adheres to AASM guidelines and includes 13 data channels (six EEG, EOG, EMG at the chin, respiratory at chest and abdomen, ECG, SaO_2, and airflow) annotated with various sleep stages and arousal categories. Annotated events include target arousals (RERA: Respiratory Effort-Related Arousals) and non-target arousals (Hypopnea, Central Apnea, Mixed Apnea, and Obstructive Apnea).

We now explore the shift from full event detection (FED) to interval-based onset detection (IOD) and consider point-based onset detection (POD) as an alternative. We utilize the DeepSleep approach for con-

tinuous segmentation (CS) and evaluate with both the traditional pointwise evaluation (PE) scheme and our ALPEC framework. The 2018 PhysioNet Challenge Dataset provides a basis for training and comparing a full event detection (FED) baseline due to its arousal annotations with both meaningful start and end points. We randomly partitioned the 994 samples into a training set of 795 samples and a test set of 199 samples. Utilizing all 13 channels, training proceeded until either early stopping criteria were met or 50 epochs were completed. The results are shown in Table 2.

Our results indicate that point-based onset detection (POD) using the DeepSleep approach for continuous segmentation is infeasible. Due to the sparsity of labels and noise in the onset annotations, the model is unable to relate meaningful patterns to arousal onsets. In pointwise evaluation (PE), a very low decision threshold results in high recall but low precision, leading to the best possible F2 score, which remains close to zero. ALPEC offers a more realistic assessment of the model's performance by discarding excessively long predicted intervals. However, if a model were to make point-predictions for arousal onsets within the temporal tolerance buffer of ground-truth onset points, ALPEC is expected to perform adequately, whereas pointwise evaluation would penalize predictions that are off by even one point. Thus, ALPEC enables the utilization and appropriate evaluation of other point-based detection approaches that have not been used before for arousal detection, such as methods for changepoint detection in time series (Aminikhanghahi and Cook, 2017).

Moreover, interval-based onset detection (IOD) performs comparably to the FED baseline when measured by ALPEC, demonstrating that detecting arousal onsets rather than full events can be equally effective. Conversely, pointwise evaluation asserts a substantial performance discrepancy, underscoring the importance of choosing an appropriate evaluation framework. An investigation of the large confidence intervals in Table 2 is deferred to Appendix B.

4.2. Experiments on the CPS dataset

The CPS dataset (see also *Data and Code Availability*) comprises 113 diagnostic polysomnographic recordings, encompassing up to 36 raw and 23 derived data channels, alongside 81 types of annotated events and additional questionnaire data for each subject. The dataset annotates various arousal classes, including those related to respiratory efforts, flow limitations, oxygen desaturation, limb movements, and spontaneous arousals. We combine all classes into a single category for binary event detection. Further details on the dataset are available in Appendix H.

Using the CPS dataset, we trained DeepSleep model candidates D1-D4, which utilize continuous segmentation on interval-based onset detection (IOD, see Section 3.1). Details about the selection and channels of the model candidates, as well as the choice of hyperparameters, are provided in Appendix D. These models are compared to popular time series classification models and naive baselines from the sktime library (Löning et al., 2019) using window-based onset detection (WOD, see Section 3.3). All models are trained using combined training and validation folds (93 subjects total). The DeepSleep models are trained until early stopping or up to 100 epochs. All models are evaluated on the held-out test set (14 subjects) using both window-based evaluation (WE) and ALPEC. All experiments are repeated five times using different random seeds. Details on data preprocessing, and data folds are available in Appendices C, and I. The results are shown in Table 3.

Both Window-Based Evaluation (WE) and ALPEC demonstrate similar performance across our interval-based onset (IOD) detection models (D1-D4), which utilize the DeepSleep approach. This consistency suggests that the domain-specific adaptations inherent in ALPEC do not drastically alter results compared to WE in this context. However, WE appears to underestimate precision, resulting in a higher count of false positives. This discrepancy is due to ALPEC's methodology of treating all adjacent points within an overlapping interval as a single true positive, unlike WE, which evaluates each window individually. Furthermore, ALPEC's buffer zones typically reduce false positives and negatives, enhancing its accuracy. In the analysis of Window-Based Onset Detection (WOD) models, WE notably overestimates recall compared to ALPEC. A clear example is observed with the *Constant 1* baseline, where WE significantly overrates its performance because this model predicts an arousal event in every window. This outcome reveals a bias in WE towards models that predict frequent events. Conversely, ALPEC shows zero performance for this baseline, effectively highlighting its ability to address the methodological shortcomings of its closest predecessor, the segment-wise f-score, as discussed in Section 3.2.

Table 2: **Comparison of modeling and evaluation approaches on the 2018 PhysioNet Challenge Dataset**. Arousal types are explained in Section 4.1. Training methods: *FED* (Full Event Detection), *POD* (Point-based Onset Detection), *IOD* (Interval-based Onset Detection) use DeepSleep (Section 3.1). Evaluation: *PE* (Pointwise Evaluation), *ALPEC* (Approximate Localization and Precise Event Count), see Sections 3.3 and 3.2. Metrics are mean values over test subjects with cross-subject validation. Models are trained five times, results averaged with 95% confidence intervals, assuming t-distributed mean values. POD is ineffective with DeepSleep; ALPEC, shows that IOD performs comparable to the FED baseline.

Arousals	Training approach	PE (baseline)			ALPEC evaluation		
		Precision	Recall	F2	Precision	Recall	F2
Target	IOD (our)	0.13 (4)	0.47 (7)	0.30 (3)	0.20 (6)	0.63 (12)	0.42 (4)
	POD (naive baseline)	7e-6	0.94	3.5e-5	0.00	0.00	0.00
	FED (baseline)	0.17 (5)	0.49 (17)	0.35 (8)	0.23 (15)	0.59 (20)	0.41 (6)
Non-target	IOD (our)	0.33 (6)	0.71 (10)	0.57 (3)	0.53 (6)	0.85 (6)	0.76 (3)
	POD (naive baseline)	3.4e-5	1.00	1.67e-4	0.00	0.00	0.00
	FED (baseline)	0.54 (4)	0.67 (8)	0.64 (5)	0.60 (4)	0.77 (5)	0.73 (2)

Table 3: **Comparison of modeling and evaluation approaches on our CPS dataset.** Metrics are presented as mean values over distinct test subjects with 95% confidence intervals, assuming t-distributed mean values, calculated over five training iterations. For window-based onset detection (WOD) and baseline models, methods from the sktime library (Löning et al., 2019), including dummy models, were used. Models D1-D4, derived from the DeepSleep approach utilizing interval-based onset detection (IOD), indicate the used channels. ALPEC offers a stringent assessment of performance, highlighting that WE tends to overestimate the effectiveness of WOD models.

	Model	Window-based evaluation (WE)			ALPEC evaluation		
		Precision	Recall	F2	Precision	Recall	F2
IOD	D4: most channels	0.49 (6)	0.82 (8)	**0.71** (3)	0.59 (8)	0.81 (8)	**0.73** (3)
	D3: no EEG, EOG, EMG	0.39 (3)	0.71 (3)	0.59 (3)	0.48 (4)	0.70 (3)	0.62 (3)
	D2: C3:A2, EOGl, EMG	0.44 (5)	0.75 (7)	0.64 (3)	0.53 (7)	0.74 (7)	0.67 (3)
	D1: C3:A2	0.40 (5)	0.76 (6)	0.62 (2)	0.48 (7)	0.75 (6)	0.65 (2)
WOD	IndividualBOSS	0.25 (0)	0.55 (1)	0.42 (1)	0.30 (1)	0.32 (2)	0.31 (2)
	SupervisedTimeSeriesForest	0.37 (0)	0.71 (1)	0.59 (1)	0.37 (1)	0.29 (1)	0.30 (1)
	TimeSeriesForestClassifier	0.30 (1)	0.65 (1)	0.50 (1)	0.30 (1)	0.30 (1)	0.29 (1)
	SignatureClassifier	0.28 (0)	0.65 (2)	0.49 (1)	0.29 (1)	0.30 (1)	0.29 (1)
	SummaryClassifier	0.27 (0)	0.62 (1)	0.48 (0)	0.27 (1)	0.29 (1)	0.28 (1)
	Catch22Classifier	0.33 (0)	0.73 (1)	0.57 (0)	0.30 (0)	0.23 (1)	0.24 (1)
Baseline	RandomStratified	0.20 (1)	0.49 (2)	0.37 (1)	0.29 (2)	0.36 (4)	0.34 (3)
	RandomUniform	0.20 (0)	0.51 (2)	0.37 (1)	0.28 (2)	0.36 (3)	0.33 (3)
	Constant 1	0.20	1.00	0.53	0.00	0.00	0.00
	Constant 0	0.00	0.00	0.00	0.00	0.00	0.00

Overall, ALPEC provides a more accurate assessment of model performance, revealing that none of the WOD models substantially outperform the random baselines. This aligns with the understanding that arousal detection is a challenging task that may not be adequately addressed by simpler classical mod-

els without specialized feature engineering (Zan and Yildiz, 2023).

Finally, we find a significantly enhanced predictive performance of model D4, which incorporates the most data modalities, compared to models D1 and D2, which use fewer modalities. Also, model D3, which does not use any electrode-based modalities, demonstrates potential for reduced technical complexity while maintaining reasonable performance.

5. Discussion

Our findings demonstrate that arousal onset detection can be effectively achieved using continuous segmentation approaches with our proposed interval-based onset detection (IOD) training scheme, achieving comparable performance to the full event detection (FED) baseline, successfully aligning model training with clinical annotation constraints for arousal annotations.

Additionally, the results highlight the significant benefits of incorporating novel data modalities, which also offer potential for reducing technical complexity. Minimizing dependence on electrode-based modalities could address issues such as electrode displacement or noise, potentially enabling home-based arousal diagnostics (Imtiaz, 2021).

A significant contribution of our work is the development of the ALPEC framework, the first performance evaluation framework tailored to the clinical requirements of arousal detection. We demonstrate that ALPEC provides a more accurate assessment of model performance compared to traditional window-based evaluation (WE) or pointwise evaluation (PE). Moreover, due to sampling at the subject-level, ALPEC overcomes common pitfalls of window-based evaluation such as class imbalance and cross-subject validation issues (see Appendix G for further details). We emphasize that our critique is not directed against window-based *classification* approaches, which remain valuable and effective in arousal detection, as shown in recent studies (Badiei et al., 2023; Foroughi et al., 2023). Our concerns specifically relate to window-based *evaluation* methods.

We advocate for the adoption of the ALPEC framework, which is immune to common pitfalls, finely tunable, and compatible with both window-based classification and continuous segmentation. ALPEC's design inherently incorporates a *precise event count* requirement while offering flexibility in *approximate location* through adjustable parameters for buffer size and maximum interval length. Additionally, it is suitable to evaluate models trained on various ground-truth annotations, including point annotations (POD), constructed intervals (IOD), and events with start and end values (FED). This adaptability makes ALPEC a versatile tool for any task requiring precise event count detection in time series data, providing a robust framework suitable for a wide range of applications in healthcare and beyond.

5.1. Limitations

(1) This work addresses binary detection of arousal onsets and does not encompass the causal differentiation of arousals, an additional task in clinical settings that necessitates a multi-class classification approach.

(2) Furthermore, ALPEC is designed solely for post-processing and performance evaluation, and does not influence the learning process of the models. While we adapted the DeepSleep method for arousal onset detection, it was not initially designed to meet the specific demands of real-world clinical applications. Future research should leverage ALPEC for comparative analysis and enhance model functionality by incorporating factors crucial for clinical decision support, such as explainability.

(3) Additionally, the settings for ALPEC's buffer size and maximum interval length hyperparameters need experimental validation in collaboration with clinical end-users to ensure their effectiveness and applicability in real-world settings.

We will address limitations two and three through an application-grounded user study in future work.

5.2. Conclusion

Our work establishes foundational elements for developing clinical decision support systems for arousal detection in sleep laboratories, addressing critical misalignments between current Machine Learning methodologies and clinical practices. We introduce the Comprehensive Polysomnography (CPS) dataset as a significant resource for sleep medical research, demonstrating the potential of utilizing novel data modalities.

Our findings contribute to the development of production-ready arousal detection models that align with current clinical annotation practices. We look forward to seeing how the research community builds on our findings and continues to evolve the field.

References

Mubashara Akhtar, Omar Benjelloun, Costanza Conforti, Joan Giner-Miguelez, Nitisha Jain, Michael Kuchnik, Quentin Lhoest, Pierre Marcenac, Manil Maskey, Peter Mattson, Luis Oala, Pierre Ruyssen, Rajat Shinde, Elena Simperl, Goeffry Thomas, Slava Tykhonov, Joaquin Vanschoren, Steffen Vogler, and Carole-Jean Wu. Croissant: A Metadata Format for ML-Ready Datasets, 2024.

Samaneh Aminikhanghahi and Diane J Cook. A survey of methods for time series change point detection. *Knowledge and information systems*, 51(2):339–367, 2017.

Adriana Anido-Alonso and Diego Alvarez-Estevez. Decentralized data-privacy preserving deep-learning approaches for enhancing inter-database generalization in automatic sleep staging. *IEEE Journal of Biomedical and Health Informatics*, 2023.

Jerome Argod, Jean-Louis Pepin, and Patrick Levy. Differentiating obstructive and central sleep respiratory events through pulse transit time. *American journal of respiratory and critical care medicine*, 158(6):1778–1783, 1998.

Afsoon Badiei, Saeed Meshgini, and Khosro Rezaee. A novel approach for sleep arousal disorder detection based on the interaction of physiological signals and metaheuristic learning. *Computational Intelligence and Neuroscience*, 2023, 2023.

Richard B Berry, Rohit Budhiraja, Daniel J Gottlieb, David Gozal, Conrad Iber, Vishesh K Kapur, Carole L Marcus, Reena Mehra, Sairam Parthasarathy, Stuart F Quan, et al. Rules for scoring respiratory events in sleep: update of the 2007 AASM manual for the scoring of sleep and associated events: deliberations of the sleep apnea definitions task force of the American Academy of Sleep Medicine. *Journal of clinical sleep medicine*, 8(5):597–619, 2012.

Bertrand. SweetViz. Visualize and compare datasets, target values and associations, with one line of code. https://github.com/fbdesignpro/sweetviz, 2020. Accessed: 2024-06-04.

Maria R Bonsignore, Tarja Saaresranta, and Renata L Riha. Sex differences in obstructive sleep apnoea. *European Respiratory Review*, 28(154), 2019.

Xiaoli Chen, Rui Wang, Phyllis Zee, Pamela L Lutsey, Sogol Javaheri, Carmela Alcántara, Chandra L Jackson, Michelle A Williams, and Susan Redline. Racial/ethnic differences in sleep disturbances: the Multi-Ethnic Study of Atherosclerosis (MESA). *Sleep*, 38(6):877–888, 2015.

Franz Ehrlich, Tony Sehr, Moritz Brandt, Martin Schmidt, Hagen Malberg, Martin Sedlmayr, and Miriam Goldammer. State-of-the-art sleep arousal detection evaluated on a comprehensive clinical dataset. *Scientific Reports*, 14(1):16239, 2024.

Ahmad Fawzy, Danastri Cantya Nirmala, Denaya Khansa, and Yudhistira Tri Wardhana. Ethics and Regulation for Artificial Intelligence in Healthcare: Empowering Clinicians to Ensure Equitable and High-Quality Care. *International Journal of Medical Science and Clinical Research Studies*, 3(07):1350–1357, 2023.

Ingo Fietze, Naima Laharnar, Anne Obst, Ralf Ewert, Stephan B Felix, Carmen Garcia, Sven Gläser, Martin Glos, Carsten Oliver Schmidt, Beate Stubbe, et al. Prevalence and association analysis of obstructive sleep apnea with gender and age differences–Results of SHIP-Trend. *Journal of sleep research*, 28(5):e12770, 2019.

Robert Fonod. DeepSleep 2.0: automated sleep arousal segmentation via deep learning. *AI*, 3(1):164–179, 2022.

Andia Foroughi, Fardad Farokhi, Fereidoun Nowshiravan Rahatabad, and Alireza Kashaninia. Deep convolutional architecture-based hybrid learning for sleep arousal events detection through single-lead EEG signals. *Brain and Behavior*, 13(6):e3028, 2023.

Karl A Franklin and Eva Lindberg. Obstructive sleep apnea is a common disorder in the population—a review on the epidemiology of sleep apnea. *Journal of thoracic disease*, 7(8):1311, 2015.

Azul Garza and Max Mergenthaler-Canseco. TimeGPT-1. *arXiv preprint arXiv:2310.03589*, 2023.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92, 2021.

Mohammad M Ghassemi, Benjamin E Moody, Li-Wei H Lehman, Christopher Song, Qiao Li, Haoqi Sun, Roger G Mark, M Brandon Westover, and Gari D Clifford. You snooze, you win: the physionet/computing in cardiology challenge 2018. In *2018 Computing in Cardiology Conference (CinC)*, volume 45, pages 1–4. IEEE, 2018.

Ary L Goldberger, Luis AN Amaral, Leon Glass, Jeffrey M Hausdorff, Plamen Ch Ivanov, Roger G Mark, Joseph E Mietus, George B Moody, Chung-Kang Peng, and H Eugene Stanley. PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. *circulation*, 101(23): e215–e220, 2000.

Matthew Howe-Patterson, Bahareh Pourbabaee, and Frederic Benard. Automated detection of sleep arousals from polysomnography data using a dense convolutional neural network. In *2018 Computing in Cardiology Conference (CinC)*, volume 45, pages 1–4. IEEE, 2018.

Kyle Hundman, Valentino Constantinou, Christopher Laporte, Ian Colwell, and Tom Soderstrom. Detecting spacecraft anomalies using lstms and nonparametric dynamic thresholding. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 387–395, 2018.

Syed Anas Imtiaz. A systematic review of sensing technologies for wearable sleep staging. *Sensors*, 21(5): 1562, 2021.

Shazia Jehan, Ferdinand Zizi, Seithikurippu R Pandi-Perumal, Steven Wall, Evan Auguste, Alyson K Myers, Girardin Jean-Louis, and Samy I McFarlane. Obstructive sleep apnea and obesity: implications for public health. *Sleep medicine and disorders: international journal*, 1(4), 2017.

Stefan Kraft, Andreas Theissler, Vera Wienhausen-Wilke, Philipp Walter, and Gjergji Kasneci. Comprehensive Polysomnography (CPS) Dataset: A Resource for Sleep-Related Arousal Research. PhysioNet data repository, 2024. URL https://doi.org/10.13026/sxs0-h317. Dataset.

Chih-Fan Kuo, Cheng-Yu Tsai, Wun-Hao Cheng, Wen-Hua Hs, Arnab Majumdar, Marc Stettler, Kang-Yun Lee, Yi-Chun Kuan, Po-Hao Feng, Chien-Hua Tseng, et al. Machine learning approaches for predicting sleep arousal response based on heart rate variability, oxygen saturation, and body profiles. *Digital Health*, 9: 20552076231205744, 2023.

Alexandre Lacoste, Alexandra Luccioni, Victor Schmidt, and Thomas Dandres. Quantifying the Carbon Emissions of Machine Learning. *arXiv preprint arXiv:1910.09700*, 2019.

H Lee, B Li, Y Huang, Y Chi, and S Lin. NCH sleep databank: a large collection of real-world pediatric sleep studies with longitudinal clinical data (version 3.1. 0). PhysioNet, 2021.

Harlin Lee, Boyue Li, Shelly DeForte, Mark L Splaingard, Yungui Huang, Yuejie Chi, and Simon L Linwood. A large collection of real-world pediatric sleep studies. *Scientific Data*, 9(1):421, 2022.

Haoqi Li, Qineng Cao, Yizhou Zhong, and Yun Pan. Sleep arousal detection using end-to-end deep learning method based on multi-physiological signals. In *2018 computing in cardiology conference (CinC)*, volume 45, pages 1–4. IEEE, 2018.

Hongyang Li and Yuanfang Guan. DeepSleep convolutional neural network allows accurate and fast detection of sleep arousal. *Communications biology*, 4(1):18, 2021.

Markus Löning, Anthony Bagnall, Sajaysurya Ganesh, Viktor Kazakov, Jason Lines, and Franz J Király. sktime: A unified interface for machine learning with time series. *arXiv preprint arXiv:1909.07872*, 2019.

Tambiama Madiega. Artificial intelligence act. *European Parliament: European Parliamentary Research Service*, 2021.

Farah Magrabi, Elske Ammenwerth, Jytte Brender McNair, Nicolet F De Keizer, Hannele Hyppönen, Pirkko Nykänen, Michael Rigby, Philip J Scott, Tuulikki Vehko, Zoie Shui-Yee Wong, et al. Artificial intelligence in clinical decision support: challenges for evaluating AI and practical implications. *Yearbook of medical informatics*, 28(01):128–134, 2019.

Matthew McDermott, Lasse Hyldig Hansen, Haoran Zhang, Giovanni Angelotti, and Jack Gallifant. A Closer Look at AUROC and AUPRC under Class Imbalance. *arXiv preprint arXiv:2401.06091*, 2024.

Daniel Miller, Andrew Ward, and Nicholas Bambos. Automatic sleep arousal identification from physiological waveforms using deep learning. In *2018 Computing in Cardiology Conference (CinC)*, volume 45, pages 1–4. IEEE, 2018.

Tomofumi Misaka, Yuko Niimura, Akiomi Yoshihisa, Kento Wada, Yusuke Kimishima, Tetsuro Yokokawa, Satoshi Abe, Masayoshi Oikawa, Takashi Kaneshiro, Atsushi Kobayashi, et al. Clinical impact of sleep-disordered breathing on very short-term blood pressure variability determined by pulse transit time. *Journal of Hypertension*, 38(9):1703–1711, 2020.

Huy Phan, Fernando Andreotti, Navin Cooray, Oliver Y Chén, and Maarten De Vos. SeqSleepNet: end-to-end hierarchical recurrent neural network for sequence-to-sequence automatic sleep staging. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 27 (3):400–410, 2019.

D Pitson, N Chhina, S Knijn, M Van Herwaaden, and J Stradling. Changes in pulse transit time and pulse rate as markers of arousal from sleep in normal subjects. *Clinical science (London, England: 1979)*, 87(2): 269–273, 1994.

DJ Pitson et al. Value of beat-to-beat blood pressure changes, detected by pulse transit time, in the management of the obstructive sleep apnoea/hypopnoea syndrome. *European Respiratory Journal*, 12(3):685–692, 1998.

Naresh M Punjabi. The epidemiology of adult obstructive sleep apnea. *Proceedings of the American Thoracic Society*, 5(2):136–143, 2008.

Stuart F Quan, Barbara V Howard, Conrad Iber, James P Kiley, F Javier Nieto, George T O'Connor, David M Rapoport, Susan Redline, John Robbins, Jonathan M Samet, et al. The sleep heart health study: design, rationale, and methods. *Sleep*, 20(12):1077–1085, 1997.

Winfried Randerath, Claudio L Bassetti, Maria R Bonsignore, Ramon Farre, Luigi Ferini-Strambi, Ludger Grote, Jan Hedner, Malcolm Kohler, Miguel-Angel Martinez-Garcia, Stefan Mihaicuta, et al. Challenges and perspectives in obstructive sleep apnoea: report by an ad hoc working group of the Sleep Disordered Breathing Group of the European Respiratory Society and the European Sleep Research Society. *European respiratory journal*, 52(3), 2018.

F Raschke and J Fischer. "Arousal" in der Schlafmedizin. *Somnologie*, 1(2), 1997.

Erik Scharwächter and Emmanuel Müller. Statistical evaluation of anomaly detectors for sequences. *arXiv preprint arXiv:2008.05788*, 2020.

Daniel J Schwartz and Pat Moxley. On the potential clinical relevance of the length of arousals from sleep in patients with obstructive sleep apnea. *Journal of Clinical Sleep Medicine: JCSM: Official Publication of the American Academy of Sleep Medicine*, 2(2):175–180, 2006.

Sobhan Salari Shahrbabaki, Dominik Linz, Simon Hartmann, Susan Redline, and Mathias Baumert. Sleep arousal burden is associated with long-term all-cause and cardiovascular mortality in 8001 community-dwelling older men and women. *European heart journal*, 42(21):2088–2099, 2021.

Sondre Sørbø and Massimiliano Ruocco. Navigating the metric maze: A taxonomy of evaluation metrics for anomaly detection in time series. *Data Mining and Knowledge Discovery*, pages 1–42, 2023.

Baptiste Vasey, Myura Nagendran, Bruce Campbell, David A Clifton, Gary S Collins, Spiros Denaxas, Alastair K Denniston, Livia Faes, Bart Geerts, Mudathir Ibrahim, et al. Reporting guideline for the early stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. *bmj*, 377, 2022.

Thomas-Christian Wetter, Roland Popp, Michael Arzt, and Thomas Pollmächer. *ELSEVIER ESSENTIALS Schlafmedizin: Das Wichtigste für Ärzte aller Fachrichtungen*. Elsevier Health Sciences, 2012.

Wienhausen-Wilke and Kraft. Computer-aided diagnostics of sleep-related arousals on the basis of pulse wave analyses. https://drks.de/search/en/trial/DRKS00033641, 2024. [Accessed: 2024-08-15].

Hasan Zan and Abdulnasır Yildiz. Multi-task learning for arousal and sleep stage detection using fully convolutional networks. *Journal of Neural Engineering*, 20(5): 056034, 2023.

Guo-Qiang Zhang, Licong Cui, Remo Mueller, Shiqiang Tao, Matthew Kim, Michael Rueschman, Sara Mariani, Daniel Mobley, and Susan Redline. The National Sleep Research Resource: towards a sleep data commons. *Journal of the American Medical Informatics Association*, 25(10):1351–1358, 2018.

Appendix A. Ablation studies

In this section, we present ablation studies conducted on our the CPS dataset to evaluate the impact of various hyperparameters on the performance of the DeepSleep method for arousal onset detection and the ALPEC framework for performance evaluation. The hyperparameters considered are parameters used in training (the smoothing window size w, and the interval length for interval-based onset detection (IOD) l) and the ALPEC framework (the maximum interval duration d, the minimum interval distance before merging δ, and the buffer size b). For a description of the parameters, see Table 9. We use the DeepSleep model architecture D1 from Table 3 with a univariate *C3:A2* channel. All models are trained using the same training and test split as in Section 4. Tables 4 to 8 show the tuning results, where parameter choices for all runs, if not tuned, are marked in bold, which are the same as for D1 in the main part of this paper.

Table 4: Smoothing window w

w	$\bar{\text{Precision}}$	$\bar{\text{Recall}}$	$\bar{\text{F2}}$
none	0.47	0.67	0.60
1	0.59	0.71	0.66
2	0.53	0.73	0.65
3	0.47	0.80	0.68
4	0.45	0.86	0.70
5	0.44	0.82	0.67

Table 5: Interval length for IOD l

l	$\bar{\text{Precision}}$	$\bar{\text{Recall}}$	$\bar{\text{F2}}$
2	0.39	0.78	0.62
6	0.49	0.67	0.61
10	0.45	0.80	0.68
14	0.47	0.80	0.68
20	0.42	0.81	0.66
30	0.47	0.59	0.55
60	0.36	0.50	0.45

We can see that most parameters have a moderate effect on performance metrics within the ranges tested. The most significant drop in performance occurs with low values of the maximum allowed interval distance d, which is expected since removing many events leads to a high number of false negatives. Values higher than $d = 60s$ make no significant difference to $d = 60s$, indicating that our ML approach produces reasonably short predicted intervals. We also see that smoothing (Table 4), merging of intervals (Table 7), and utilizing a buffer (Table 8) all lead to performance improvements. Smoothing and merging actually affect the predicted intervals, whereas the buffer only affects the evaluation by relaxing the locality requirement.

Table 6: Max. interval duration d

d	$\bar{\text{Precision}}$	$\bar{\text{Recall}}$	$\bar{\text{F2}}$
10	0.57	0.19	0.22
30	0.50	0.74	0.66
60	0.47	0.80	0.68
90	0.43	0.77	0.65
120	0.40	0.86	0.67
none	0.45	0.82	0.68

Table 7: Min. interval distance before merging δ

δ	$\bar{\text{Precision}}$	$\bar{\text{Recall}}$	$\bar{\text{F2}}$
0	0.48	0.76	0.66
5	0.50	0.74	0.66
10	0.47	0.80	0.68
15	0.37	0.86	0.66
20	0.37	0.58	0.50

Table 8: Buffer size b with $b = b^{\text{before}} = b^{\text{after}}$

b	$\bar{\text{Precision}}$	$\bar{\text{Recall}}$	$\bar{\text{F2}}$
0	0.51	0.69	0.63
5	0.47	0.70	0.61
10	0.46	0.75	0.65
15	0.47	0.80	0.68
20	0.43	0.82	0.68
25	0.59	0.66	0.63

Appendix B. Analysis of the effects of decision thresholds

The rather large confidence intervals in the results on the 2018 PhysioNet Challenge Dataset (Section 4,

Table 9: Table of notation

Symbol	Meaning
$\mathbf{x}$	Multivariate input sequence
n	Number of data points contained within each input channel after padding with zeros, fixed to 2^{23}
D	Dataset containing all subjects
T	Training set containing a subset of subjects
V	Validation set containing a subset of subjects
$p_i(\mathbf{x})$	Probability score of the i-th time step in the input sequence being of the positive class
$p_\eta(\mathbf{x})$	Probability score for the η-th window in the input sequence
$c_i(\mathbf{x})$	Binary class prediction for the i-th time step
N	Number of windows when splitting the input sequence into windows of length s
s	Length of each window when splitting the input sequence into N windows
w	Window size for smoothing the probability scores
f	Sampling frequency of the data
t_k	Threshold for converting probability scores to binary class predictions. We use 101 thresholds from 0 to 1 in steps of 0.01, i.e. $k = 0, 1, ..., 100$
δ	Minimum distance in seconds for merging two adjacent predicted intervals in ALPEC
C	Binary class predictions for the whole input sequence, i.e. $C = \{c_1, c_2, ..., c_n\}$
I	Predicted interval in binary class predictions C with start and end indices I^{start} and I^{end}
G	Ground-truth interval with start and end indices G^{start} and G^{end}
d	Maximum duration of a predicted interval before its removal in ALPEC
P	Predicted interval with start and end indices P^{start} and P^{end}
b^{before}	Temporal tolerance buffer before the ground-truth interval in ALPEC
b^{after}	Temporal tolerance buffer after the ground-truth interval in ALPEC
G^{ext}	Extended ground-truth interval with start and end indices $G^{\text{start,ext}}$ and $G^{\text{end,ext}}$
l	Length of the interval around an onset point for interval-based onset detection (IOD)

Table 2) stem from the variance in selected decision thresholds for individual runs, as shown in Table 10.

Table 10: **Comparison of selected decision thresholds (DT) and their effects**. This table presents the decision thresholds for five individual runs with different random seeds, leading to the results for the *Target* arousals using *FED* training and *ALPEC* evaluation shown in Table 2.

Run	DT	$\overline{\text{Precision}}$	$\overline{\text{Recall}}$	$\overline{\text{F1}}$	$\overline{\text{F2}}$
1	0.22	0.44	0.42	0.43	0.43
2	0.03	0.17	0.82	0.47	0.47
3	0.07	0.18	0.68	0.44	0.44
4	0.13	0.19	0.48	0.37	0.37
5	0.12	0.15	0.53	0.35	0.35

These thresholds are based on samples from the training fold and are automatically selected to maximize the F2 score. Comparing runs 1 and 2, for example, shows that a lower decision threshold results in higher recall but lower precision, as expected, while the resulting F2 scores are comparable.

Appendix C. Data preprocessing

For our experiments, all raw data channels undergo third-order Butterworth bandpass filtering to remove noise. Critical frequencies for the Butterworth bandpass filter for the different data modalities are listed in Table 11.

The raw data channels are then normalized using z-score normalization. Derived channels are upsampled to 256 Hz using repeated values and scaled to a range of $[0, 1]$ via min-max normalization. Channels are padded symmetrically to a fixed length of $n = 2^{23}$, or approximately 9 hours, to accommodate the longest recording. Magnitude scaling is applied randomly between 0.8 and 1.25 during training to enhance model generalization Li and Guan (2021).

All nominal event data are encoded as binary features, with each event type represented as a separate feature. For Sleep Profile and Body Position events, we utilize a one-hot encoding scheme to represent the different classes.

Appendix D. Hyperparameter tuning details

In this section, we provide an overview of the selected hyperparameters and perform preliminary experiments with the DeepSleep approach for continuous segmentation on unimodal data channels to find a good set of input channels for final model candidates. All models are trained on the training set and evaluated on a fixed validation set. See Appendix I for details on the data splits. Table 12 contains an overview of the hyperparameters used in this work.

Selection of input channels We perform two baseline sets of tuning runs: One on raw channels (see Table 15) and another on derived channels and a promising selection of event channels (see Tables 16 and 17). From the raw channels, we leave out the *Battery* and *REM Confidence* channels since we expect those to be irrelevant for arousal detection. Also, we only use one channel each from the EEG, EMG, and EOG groups. We split the categorical event channels *Sleep Profile* and *Body Position* into singular channels using a one-hot encoded representation of the categories. For simplicity, we will refer to both the derived and event channels as *derived* channels from here on. All results from these runs are shown in Tables 13 and 14.

Tuning all possible combinations of raw and derived channels from Tables 15, 16, and 17 would be computationally very demanding, even when restricting ourselves to the most discriminative channels. From explorative experiments, we learned that the performance of DeepSleep generally increases when using additional channels. Therefore, we selected four combinations of channels as model candidates for the final evaluation in the main part of this work, denoted with a *D* for *DeepSleep*:

1. D1, using only the channel *C3:A2*, which yielded the best performance in Table 13 and is the required choice for manual arousal detection according to the AASM guidelines (Berry et al., 2012).

2. D2, using modalities often selected for arousal detection in related work (see Table 1), namely *C3:A2*, *EOGl*, and *EMG*.

3. D3, using a selection of channels that do not rely on EEG, EMG, and EOG modalities as indicated in Tables 13 and 14 in the *D3* column.

Table 11: Critical frequencies for the bandpass filter for different modalities

Modality	**Channels**	**Lower freq. [Hz]**	**Upper freq. [Hz]**
EEG	C4:A1, C3:A2, F4:A1, O2:A1, A1, A2, C3, C4, F4, O2	0.2	35
EOG	EOGl, EOGL:A1, EOGL:A2, EOGr, EOGr:A1, EOGr:A2	0.2	35
EMG	EMG+, EMG-, EMG	10	127
ECG	ECG 2	0.2	127
Respiratory	Pressure Flow, Thermal Flow	0.001	15
Snore	Snoring Pressure, Snoring Sound	20	127
PPG	Pleth	0.5	5

Table 12: **Choices for Hyperparameters**. Values in seconds are multiplied by the fixed sampling rate of 256 Hz. For further explanations of the meaning of the symbols, see Table 9

Context	Parameter	Value	Explanation
Data (CPS)	$\|T\|$	64	Number of subjects in the training set
	$\|V\|$	28	Number of subjects in the validation set
	$\|E\|$	14	Number of subjects in the test set
Training	n	2^{23}	Number of padded data points per channel
	s	$30s$	Window size for window-based classification
	ω	3s	Smoothing window for continuous segmentation
	l	10s	Interval length for IOD
	epochs	100	Maximum number of training epochs
	batch size	1	Number of subjects per batch
ALPEC	d	60s	Maximum interval duration
	δ	10s	Minimum interval distance
	b^{before}	15s	Left buffer size
	b^{after}	15s	Right buffer size

Table 13: Unimodal training on raw data channels. *D3* and *D4* are selections of channels that are used as model candidates in the main part of this paper.

Channel	$\bar{F}_2$	D3	D4
C3:A2	0.60		✓
Pressure Flow	0.56	✓	✓
RIP.Abdom	0.53	✓	✓
EMG	0.53		✓
Sum RIPs	0.52	✓	✓
EOGl	0.50		✓
Pulse	0.48	✓	✓
RIP.Thrx	0.46	✓	✓
Pleth	0.45	✓	✓
Snoring Pressure	0.45	✓	✓
Thermal Flow	0.43	✓	✓
PLMl	0.37	✓	✓
ECG 2	0.36	✓	✓
Light	0.34	✓	✓
SPO2	0.33	✓	✓
Snoring Sound	0.30	✓	✓
Motion	0.28	✓	✓

4. D4, using all channels from Tables 13 and 14 except for the most underperforming channels, *Body Position* and *Central Apnea*, also indicated in the *D4* column of the tables.

Additionally, we did not select the *Heart Rate*, *Light*, and *SpO2* derived channels for model candidates *D3* and *D4* since these are very similar to the *Pulse*, *Light*, and *SPO2* raw channels, respectively, as indicated by the similar performances in Tables 13 and 14.

The effects of additional hyperparameters are detailed in Appendix A, where calculations are performed using the *D1* unimodal channel selection.

Appendix E. Schematic comparison of training and evaluation schemes

In this work, we utilize a multitude of training and evaluation approaches which are schematically illustrated in Figure 2. For an explanation of the schemes, we refer to Section 3 and Appendix F.

We now want to perform a more detailed conceptual comparison of our proposed Approximate Lo-

Table 14: Unimodal training on derived channels. *D3* and *D4* are selections of channels that are used as model candidates in the main part of this paper.

Channel	$\bar{F}_2$	D3	D4
Average Frequency Value	0.55		✓
Hypopnea	0.53	✓	✓
Sigma FFT	0.50		✓
Heart Rate	0.48		
RR Interval	0.45	✓	✓
Delta FFT	0.45		✓
Alpha+Beta FFT	0.44		✓
PTT Raw	0.44	✓	✓
HRV LF	0.43	✓	✓
Diastol	0.43	✓	✓
Obstruction	0.43	✓	✓
Systol PTT	0.42	✓	✓
Sleep Profile	0.42		✓
Syst	0.42	✓	✓
Diastol PTT	0.41	✓	✓
RR	0.41	✓	✓
SVB	0.41	✓	✓
Phase Angle	0.41	✓	✓
Light	0.33		
SpO2	0.31		
Activity	0.28	✓	✓
Integral EMG	0.26		✓
HRV HF	0.25	✓	✓
Obstructive Apnea	0.17	✓	✓
Apnea	0.09	✓	✓
Body Position	0.04		
Central Apnea	0.04		

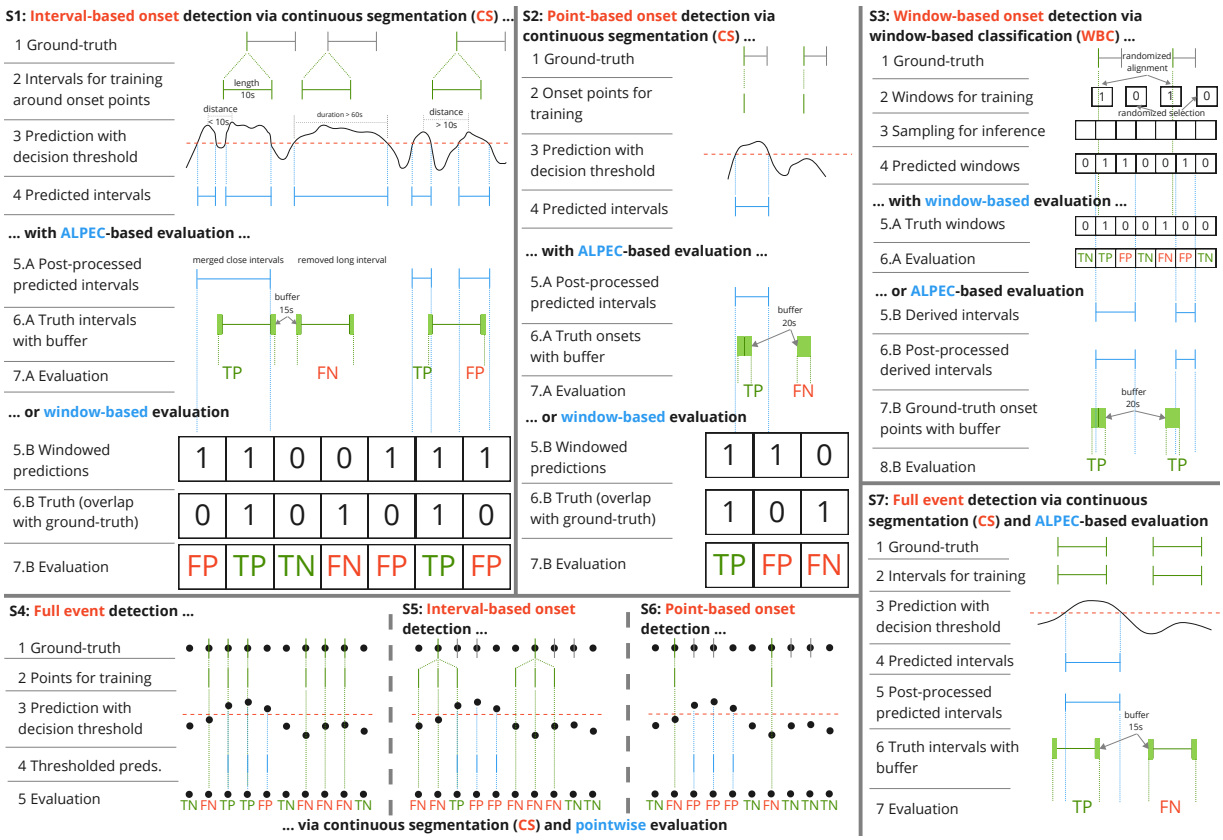


Figure 2: **Schematic illustration of different approaches for training and evaluating arousal detection models.** In schemas S1, S2, S3, and S7, lines and areas in green color represent target points of the positive class (arousal) while empty areas in between contain points of the negative class (no arousal). Lines in blue color represent points that are predicted to be in the positive class. For schemas containing pointwise evaluations (S4-S6), all points which are marked in green or blue are considered to be in the positive class while all other points are considered to be in the negative class. Names of training schemes are highlighted in red, evaluation schemes in blue. All sizes and dimensions are for illustrative purposes and not representative. Especially schemas containing pointwise evaluations will contain many more data points inside events/intervals (S4-S5) and between events (S4-S6). For schemas containing window-based approaches (S1-S3), each box represents a window of fixed length containing many data points, where the class identification or evaluation outcome of each point is given by the label on the box.

calization and Precise Event Count (ALPEC) framework with the baselines of pointwise evaluation and window-based evaluation. We start by remembering that, from a clinical productional point-of-view, the most important aspect of arousal diagnostics is to detect the correct number of arousals and to locate them approximately correctly to enable human validation (cf. Section 3.2). Looking at the *Evaluation* steps of S4-S6 in Figure 2, we see that pointwise evaluation is inadequate based on our requirements since it sanctions every wrong prediction point. This leads to a bias towards favoring models that strictly predict the exact labeled points which might also result in overfitting and a lack of the generalization capabilities of models that are optimized under pointwise evaluation. When comparing the exact situation with ALPEC instead of pointwise evaluation (S7), we see that ALPEC is not concerned with single points but only close-by intervals and counts exactly one TP and one FN as would be expected in this situation from a clinical point-of-view. As Sørbø and Ruocco (2023) have noted generally about pointwise evaluation, a major shortcoming is the lack of tolerance which renders it inappropriate for the evaluation of arousal detection models.

Moving on to window-based evaluation, we find a similar situation as with pointwise evaluation although less severe. Looking at S3, we see that window-based evaluation sanctions the third predicted window although adjacent to a correct prediction (the second) and sanctions the second arousal twice (one FN, one FP), for the close miss of the onset point at the border to the window with a predicted 1. To be fair, window-based evaluation could be equipped with similar domain-specific adaptations like the interval-based ALPEC. Our fundamental critique, however, is that with a windowing approach there are always technical constraints due to the window size which lead to deviations from the intended goal which can be utterly avoided by using the interval-based approach of ALPEC. It entails intrinsic flexibility, allowing it to be closely adapted to the clinical needs. Apart from this, as we have seen in Section 2.1, window-based classification approaches often contain many hyperparameters related to window size, overlap, and voting strategies which often extend to window-based evaluation. ALPEC contains hyperparameters of its own which, however, are less technical and instead are introduced to foster an adaptation to the productive clinical requirements.

Appendix F. Algorithmic description of ALPEC

Algorithm 1 provides an algorithmic description of ALPEC.

ALPEC is compatible with both window-based classification (WBC) and continuous segmentation (CS) approaches to arousal detection. For WBC, we process either probability scores $p_\eta(\mathbf{x})$ or binary predictions c_η for each window η. For CS, the process begins with probability scores $p_i(\mathbf{x})$ for each data point i. In cases where probability scores are available, the optimal threshold t_{opt} is determined from the training set T to convert these scores into binary predictions (Algorithm 1, line 2). For each subject ν and threshold t, we post-process the predictions by applying thresholding and resampling for WBC (lines 17 and 18) or just thresholding for CS (line 20). Next, predicted intervals that are less than $\delta = 10$ seconds apart are merged (line 21). For full event detection, merging is based on the closest points of predicted intervals, while for arousal onset detection, it is based on the maxima of the prediction scores, indicating the most likely points of arousal onset.

After post-processing, predictions are compared to ground-truth data G – which may consist of full event annotations, point annotations, or constructed intervals (see Section 3.1) – to determine true positive (TP), false positive (FP), and false negative (FN) counts (line 12).

ALPEC introduces two key *approximate localization* components. First, a temporal tolerance buffer (Scharwächter and Müller, 2020) of 15 seconds is applied before (b^{before}) and after (b^{after}) each ground-truth interval (line 25). Predicted intervals overlapping with ground-truth intervals within this buffer are counted as TPs. Second, ALPEC restricts the maximum duration of predicted intervals, with only those shorter than $d = 60$ seconds qualifying as TPs (line 27).

The counting method in ALPEC (lines 26 to 32) fulfills the *precise event count* requirement. A TP is recorded when any eligible predicted interval P overlaps with a buffered ground-truth interval G. A FN is recorded if a G does not overlap with any P, and a FP is noted if a P does not overlap with any G. Overly long predicted intervals contribute to only one TP, and multiple ground-truth intervals spanned by a single predicted interval result in multiple FNs unless each ground-truth interval is uniquely matched to a predicted interval.

Algorithm 1 ALPEC post-processing and performance evaluation framework. This compact representation assumes data and main input as globally accessible. The *Eval* function is a placeholder for known implementations in the literature to calculate metrics from TP, FP and FN counts. For-loop variables used outside their scope imply storage in accumulative data structures.

Data: Multivariate input channels $\mathbf{x}$, training set T, validation set V, ground-truth intervals G

Input: Probability scores $p_i(\mathbf{x})$ or $p_\eta(\mathbf{x})$ for each data point i or window η or binary predictions c_η for each window, and hyperparameters: Minimum interval merge distance δ, maximum predicted interval duration d, ground-truth temporal tolerance buffers b^{before} and b^{after}

Output: Mean values for precision, recall and F2-score over subjects in V

1 **if** *Input contains probability scores* $p_i(\mathbf{x})$ *or* $p_\eta(\mathbf{x})$ **then**
2 $t_{\text{opt}} \leftarrow$ `DetermineOptimalThresholdOnTrainingSet()` ; `// Get optimal threshold`
3 **else** `//` Input contains binary predictions c_η
4 $t_{\text{opt}} \leftarrow None$; `// No thresholding`
5 **foreach** *subject* ν *in* V **do**
6 $\text{precision}_\nu, \text{recall}_\nu, \text{F2}_\nu \leftarrow$ `Eval(CompareTruthPredPerSubject(PostProcPreds(`ν, t_{opt}`)))`
7 $\bar{\text{precision}}, \bar{\text{recall}}, \bar{\text{F2}} \leftarrow$ Compute mean values over subjects ν ; `// Get` $\bar{\text{precision}}, \bar{\text{recall}}, \bar{\text{F2}}$
8 **Function** `DetermineOptimalThresholdOnTrainingSet():`
9 **foreach** *subject* ν *in* T **do**
10 **for** *threshold* $t = 0, ..., 1$ *in steps of* 0.01 **do**
11 $c_{i\nu} \leftarrow$ `PostProcPreds(`ν, t`)` ; `// Post-processing`
12 $\text{F2}_{\nu t} \leftarrow$ `Eval(CompareTruthPredPerSubject(`$c_{i\nu}$`))` ; `// Compare intervals`
13 $t_{\text{opt}} \leftarrow$ Get threshold t with the highest average $\text{F2}_{\nu t}$ over ν ; `// Find optimal F2`
14 **return** t_{opt} ; `// Return optimal threshold`
15 **Function** `PostProcPreds(`ν, t`):`
16 **if** *window-based classification* **then**
17 Convert $p_{\eta\nu}(\mathbf{x})$ to binary predictions $c_{\eta\nu}$ if $t \neq None$; `// Thresholding`
18 Resample $c_{\eta\nu}$ to get binary predictions $c_{i\nu}$ per data point ; `// Resampling`
19 **else** `//` Continuous segmentation
20 Convert $p_{i\nu}(\mathbf{x})$ to binary predictions $c_{i\nu}$ using threshold t ; `// Thresholding`
21 Merge intervals in $c_{i\nu}$ closer than δ ; `// Interval merging`
22 **return** $c_{i\nu}$; `// Return post-processed predictions`
23 **Function** `CompareTruthPredPerSubject(`c_i`):`
24 Init TP, FP, FN to zero and empty set M^{P} for tracking matched predicted intervals P in c_i;
25 Extend each true interval G by b^{before} and b^{after} ; `// Buffer ground-truth`
26 **foreach** *extended ground-truth interval* G **do**
27 **if** *at least one overlap of* G *with any* $P \notin M^P$ *exists with length(*P*)*$\leq d$ **then** `// Selecting`
28 Add first overlapping P to M^{P}; `// Track matched interval`
29 TP++ ; `// Increment TP`
30 **else**
31 FN++ ; `// Increment FN`
32 Set FP to the number of predicted intervals P not in M^{P} ; `// Count FP`
33 **return** TP, FP, FN ; `// Return TP, FP, FN`

Selecting appropriate metrics is crucial for evaluating and analyzing model performance. Following the taxonomy by Sørbø and Ruocco (2023), we select the F2 score as the final metric for optimization (performance evaluation) and use precision and recall as auxiliary metrics for additional insights (performance analysis). ALPEC computes the micro-average F2 scores across all subjects in the training set T to determine the optimal decision threshold t_{opt} (line 13). This threshold is then used to calculate the metrics

for each subject in the validation set V, which may also be the test set (line 6). Results are aggregated using the mean to adequately represent individual outliers (line 7).

Appendix G. Overcoming evaluation pitfalls with ALPEC

Authors employing window-based evaluation often overlook reporting the class balance between arousal and non-arousal samples. In instances where the balance is disclosed, such as in the work of Kuo et al. (2023), who reported a ratio of 42,311:33,479 (arousals vs non-arousals), and Badiei et al. (2023), whose confusion matrices implied a ratio of about 1:2, discrepancies arise. Our CPS dataset indicates an expected ratio of about 1:5 for 30-second windows, based on the average total sleep time and the number of arousals across subjects. Such disparities are problematic for comparative analyses and from a production standpoint, as they likely lead to underestimations of false positives when background samples are underrepresented. Our ALPEC framework addresses this by sampling at the subject level rather than the window level, ensuring that validation samples are representative of the overall dataset.

Moreover, the lack of cross-subject validation is a frequent oversight with window-based evaluations, where samples (intervals) from all subjects are often mixed across training, validation, and test sets. Since production models are applied to unseen subjects, it is critical to evaluate these models on new subjects during development. This practice is not consistently reported, which can inflate perceived model performance. ALPEC inherently avoids this issue by enforcing subject-level sampling, ensuring that the division of training, validation, and test samples maintains subject integrity. This approach enhances the comparability of results across studies and provides a more authentic evaluation of model efficacy.

Appendix H. CPS dataset details

A detailed description of all channels and fields within the dataset, translations of data fields from German to English, and Croissant (Akhtar et al., 2024) metadata (under the Apache-2.0 license) are provided on the PhysioNet page of the CPS dataset (Kraft et al., 2024). To avoid redundancy and potential ambiguities in case of updates on the PhysioNet page, we have not included this information in this paper. All relevant information about the data used in in the main part of this paper is however described in the following.

Inclusion and exclusion criteria Patients included in the dataset were aged 18 or older and referred for polysomnographic examination at a sleep laboratory. Patients undergoing diagnostic treatments in the form of positive airway pressure therapies were excluded.

Data extraction and preprocessing The data extraction involved multiple steps using the SOMNOscreen device from SOMNOmedics GmbH, capturing a broad range of physiological signals. The data was further processed using the DOMINO software from the same manufacturer, which calculated additional data channels and provided initial annotations for sleep stages and arousals, which were manually reviewed and adjusted by medical experts from NRI Medizintechnik GmbH, according to guidelines from the American Academy of Sleep Medicine (AASM) (Berry et al., 2012). The raw data channels were upsampled to a uniform sampling rate of 256 Hz.

All input features used in this work are described in Table 15 (raw measurement data), Table 16 (derived channels), and Table 17 (nominal event data).

Additional preprocessing of the data channels before release involved shifting the day, month, and year of all recordings to January 1, 1970, to ensure patient anonymity and converting from the European Data Format (EDF) to the Waveform Database (WFDB) format.

The target arousal classes are listed in Table 17. The presence of the postfix *(EEG)* at an arousal event class indicates that the arousal was first recognized in the EEG channel, followed by its causative occurrence. In contrast, the lack of *(EEG)* denotes that the causative event preceded the observable EEG effects. Another class of arousals that is also annotated but not included in this work are autonomic arousals. This exclusion is based on the distinct nature of autonomic arousals, which involve involuntary physiological responses regulated by the autonomic nervous system, differing from arousals typically detected in sleep studies through EEG or related to specific sleep disturbances. Autonomic arousals may not directly correlate with sleep architecture changes or the specific arousal events typically analyzed in sleep medicine, thus requiring separate consideration from

Table 15: Raw data channels

Channels	**Description**
C4:A1, C3:A2, F4:A1, O2:A1, A1, A2, C3, C4, F4, O2	Electroencephalogram. Single electrodes mean that this electrode is derived against all other electrodes.
Battery	Battery voltage level
Motion	Movement sensor measuring patient's physical activity or motion
Pressure Flow	Airflow pressure measured using oxygen nasal cannula at the nose and mouth
Thermal Flow	Thermal airflow sensor measuring breathing flow rate
ECG 2	Electrocardiogram measuring heart's electrical activity
EMG+, EMG-, EMG	Electromyogram measuring skeletal muscle activity at the left side (-) and right side (+) of the chin
EOGl, EOGL:A1, EOGL:A2, EOGr, EOGr:A1, EOGr:A2	Electrooculogram measuring the left (l) and right (r) eye movements
Light	Ambient light sensor measuring light exposure
PLMl, PLMr	Periodic Limb Movement sensors measuring limb movements at the left leg (l) and right leg (r)
Pleth	Plethysmography measuring changes in blood volume at the tip of the ring finger of the non-dominant arm
Pos.	Body position sensor. Used to derive the patient's posture
Pulse	Pulse rate of the pulse wave
RIP.Abdom, RIP.Thrx, Sum RIPs	Respiratory Inductance Plethysmography sensors measuring abdominal and thoracic movements during breathing. *Sum RIPs* is a combination of *RIP.Abdomen* and *RIP.Thrx*
SPO2	Pulse oximetry sensor measuring blood oxygen saturation levels
Snoring Sound	Snore sensor measuring snoring sounds or vibrations
Snoring Pressure	Pressure sensor measuring snoring intensity using oxygen nasal cannula at the nose and mouth

Table 16: Derived signals which are calculated by the DOMINO Software from the raw data

Signal name	**Description**
Syst	Systolic blood pressure curve
Diastol	Diastolic blood pressure curve
MAP	Mean arterial pressure
Diastol PTT	Diastolic pulse transit time
Systol PTT	Systolic pulse transit time
SpO2	Average oxygen saturation level
Integrated EMG	Integrated electromyography signal from the chin
PTT Raw	Pulse transit time
HRV LF	Low frequency component of heart rate variability
HRV HF	High frequency component of heart rate variability
Heart rate	Heart rate curve
RR Interval	RR interval for heart rate analysis
SVB	Sympathovagal balance of sympathetic and parasympathetic activity
RR	Respiratory rate per minute
Obstruction	Obstruction curve in synchronized effort from abdomen and thorax
Phase Angle	Phase angle of synchronized effort
Alpha+Beta FFT	Alpha and beta wave frequency analysis in sleep
Delta FFT	Delta wave frequency analysis in sleep
Sigma FFT	Sigma wave frequency analysis in sleep
Average Frequency Value	Average frequency value in sleep FFT analysis
Activity	Activity level
Light	Light intensity in lux

Table 17: Annotated events that are used in this work. For a complete list of all annotated events, refer to the official CPS dataset documentation (Kraft et al., 2024).

Event name	**Description**
Respiratory Arousal (EEG)	EEG arousal due to respiratory effort
Respiratory Arousal	Arousal due to respiratory effort
Flow Limitation Arousal (EEG)	EEG arousal due to flow limitations
Flow Limitation Arousal	Arousal due to flow limitations
SpO2 Arousal (EEG)	EEG arousal due to oxygen desaturation
LM Arousal (EEG)	EEG arousal due to limb movements
LM Arousal	Arousal due to limb movements
PLM Arousal (EEG)	EEG arousal due to periodic limb movements
PLM Arousal	Arousal due to periodic limb movements
Snoring Arousal (EEG)	EEG arousal due to snoring
Snoring Arousal	Arousal due to snoring
Arousal (EEG)	Spontaneous EEG arousal
Arousal	Spontaneous arousal
Sleep Profile: N1	N1 sleep stage
Sleep Profile: N2	N2 sleep stage
Sleep Profile: N3	N3 sleep stage
Sleep Profile: Rem	Rapid Eye Movement sleep stage
Sleep Profile: Wach	Awake state during the measurement
Body Position: Prone	Prone body position
Body Position: Upright	Upright body position
Body Position: Left	Lying on the left side
Body Position: Right	Lying on the right side
Body Position: Supine	Supine body position
Hypopnea	Hypopnea event
Apnea	Apnea event
Central Apnea	Central apnea event
Obstructive Apnea	Obstructive apnea event

a sleep medical perspective. Autonomic arousals are also typically not included in other ML-based works which focus on the general arousal detection task.

Loading the dataset The offical documentation of the CPS dataset on PhysioNet (Kraft et al., 2024) contains code files and instructions on how to load the dataset based on Croissant specifications (Akhtar et al., 2024). These are also attached in the supplementary material. Use instructions are provided in the *README.md* file.

Dataset statistics and analysis of representativeness Table 18 provides an overview of the CPS dataset, including demographic information, sleep architecture, and sleep disorder indices.

Since the CPS dataset was collected over the course of one year during routine clinical practice, it is expected to be representative of patients undergoing polysomnographic examinations in a sleep laboratory. Our analysis reveals that close to 80% of the patients in our study suffer from obstructive sleep apnea (OSA) of varying severity. The dataset features a male-to-female ratio of 45:17 (noting that gender information is not available for all patients), with about 70% of the patients being over 50 years old, and over 40% classified as obese (BMI $>$ 30). These characteristics align with findings in existing literature, which indicate that OSA is more prevalent among males, older individuals, and those with obesity. Such demographic patterns are well-documented in research, supporting the representative nature of the CPS dataset (Bonsignore et al., 2019; Jehan et al., 2017; Fietze et al., 2019).

The official CPS dataset page on PhysioNet and the supplementary material include a script named *generate_statistics.py* that uses the data loading functions described in Section H to load the data. This script generates basic demographic statistics, statistics on questionnaire answers, medical diagnoses, and additional derived statistics (e.g. distribution of the number of arousals per subject). All statistics are automatically generated using the SweetViz (Bertrand, 2020) library, available under the MIT license. Precomputed statistics are provided in the file *statistics.html*.

Appendix I. Data folds

The split between training and validation samples is performed randomly. The test set, however, was hand-selected by an expert in arousal diagnostics whose task was to find a set of samples that is representative of patients undergoing polysomnographic examinations in a sleep laboratory. A mapping of sample IDs to folds can be found in Tables 19 to 21. A few samples are excluded in our experiments: Five entries, marked with †in the training and validation folds, since they do not contain *Diastol* and *Syst* derived channels and one entry, marked with †in the test set, since it contains overlapping target annotations.

Appendix J. Datasheet

The following datasheet is based on the *Datasheets for Datasets* framework Gebru et al. (2021).

J.1. Motivation

For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.
The CPS dataset was compiled to conduct a clinical study on arousal diagnostics (Wienhausen-Wilke and Kraft, 2024). The primary study goals are to investigate if Machine Learning can enhance the quality and efficiency of sleep-related arousal diagnostics, while also reducing technical demands. The dataset was created with the specific task of refining the diagnostic workflow by leveraging Photoplethysmography (PPG) data to reduce reliance on comprehensive EEG, electromyography (EMG), and electrooculography (EOG) inputs.

Who created the dataset (e.g., team, research group) and on behalf of which entity (e.g., company, institution, organization)?
The dataset was compiled from patients undergoing regular polysomnographic examinations at the sleep laboratory of the Klinik für Kardiologie, Pneumologie und Angiologie at Klinikum Esslingen. The companies IT-Designers Gruppe and NRI Medizintechnik GmbH were involved in collecting and processing the dataset. IT-Designers Gruppe initiated, funded, and supported the research. NRI Medizintechnik GmbH operates the sleep laboratory and cooperated and assisted in implementing the data collection protocol. Technical support was provided by SOMNOmedics GmbH, the supplier for the hardware and software of the sleep laboratory. The clinic and companies are all based in Germany.

Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grant and the grant ID.

Table 18: **Characteristics of the CPS dataset.** The total number of patients is 113. The number of patients who indicated their gender in the questionnaire was 62, no other genders were mentioned. The number of patients with REM sleep was 110.
Data are expressed as absolute values or as mean values ± standard deviation in the units provided.
OSA Severity is determined based on the AHI (Wetter et al., 2012), where *Very Severe* is an additional category reserved for patients with *Severe* OSA and additionally a high hypoxemia burden and high daytime sleepiness with a tendency to fall asleep during the day. The Baveno Classification for OSA severity (Randerath et al., 2018) was newly introduced during the collection of the CPS dataset.
Abbreviations: BMI: Body Mass Index, TST: Total sleep time; WASO: Wake time after sleep onset; Sleep Stages: N1, N2, N3: Stages of non-rapid eye movement sleep, REM: Rapid eye movement sleep; OSA: Obstructive sleep apnea; AHI: Apnea–hypopnea index; ArI: Arousal index; ESS: Epworth Sleepiness Scale; SPO_2: Oxygen saturation level; ODI: Oxygen desaturation index ($\geq 3\%$); T90: Time percentage spent below 90% oxygen saturation during sleep.

Age (years)		**OSA Severity**	
<50	26 (23.01%)	Mild	10 (8.85%)
50-60	34 (30.09%)	Moderate	39 (34.51%)
60-70	23 (20.35%)	Severe	22 (19.47%)
>70	22 (19.45%)	Very Severe	17 (15.04%)
Unknown	8 (7.08%)	Other	25 (22.12%)
BMI (kg/m^2)		**Baveno Classification**	
18.5-25	19 (16.81%)	Type A	15 (13.27%)
25-30	43 (38.05%)	Type B	32 (28.32%)
>30	48 (42.48%)	Type C	8 (7.08%)
Unknown	3 (2.65%)	Type D	1 (0.88%)
Gender		Unknown	57 (50.44%)
Male	45 (72.58%)	**Mean ESS**	7.83 ± 4.96
Female	17 (27.42%)	**Mean Number of Arousals**	167.56 ± 87.17
Sleep Architecture		**Sleep Disorder Index**	
Sleep Efficiency (%)	70.04 ± 15.87	AHI (events/hour)	107.86 ± 39.60
TST (min)	435.48 ± 36.83	ArI (events/hour)	20.95 ± 10.42
WASO (min)	130.45 ± 69.35	Snoring Index (events/hour)	65.25 ± 106.53
N1 (% of TST)	15.74 ± 11.70	**Oximetry Parameters**	
N2 (% of TST)	51.22 ± 11.58	SPO_2 (%)	93.75 ± 1.67
N3 (% of TST)	19.31 ± 11.62	ODI (events / hour)	22.23 ± 15.63
REM (% of TST)	13.74 ± 6.40	T90 (%)	8.04 ± 11.66

Table 19: Training set

3DquDEk2YwjfckxNBAQuVTshrK3VWqO7
HvVu33fnVKDLLjwY8Mtytcgi8Btsr1kS
INzmELsQB5yeF6HnHRM76U1ufVy7vmfb
hOwipKAoUqK6vJDqsjnchMKZf0e9uSH8
5Nl24yHOnojhw7nsgC0e530b2RBzOuLA
1RLVk0ocGDZLI8RRhPglAc4I3gMSLqvu
0Ah95Qw18puf1JsnrKBA6u8XXZL1MIQJ†
Bmy6KwUhfqRDp6bzRx1PaWoQvBpImF01
FddiLTFWMZFHH5s1NFddllezef4BJhwS
tfAnzkFia5hzaA6bHYFpkj3jPF9OFzAj
tU3dZpxIdmbr9wpPpFeZGh5MciOB1TgT
FPSnBoS217CEJ8cZS7MO3VuYUJwIt8LV
5wPINWASVdhb63RK4DtJt5LuyuaWyMo8
JyXyuQIuFyAtKL8ZoWS98xvpW4PJFco6
KIdPVCRCkXIaWDQ4c4gU38xpH5PAnOSV
CMTsLOEWEJvQqMeOGKKCKN87IXp9LOUU
0pdaTB9613iRUlhUJRAYvKmMQcOV3TYn
1gQ3otWoJ3qNNJ5g4N1WtTC4JTFOlP9B
wl5eUqFCFGibvn13l8axb82mNOkp0doc
DlUvX2vg1c4ORh63BRFaCxbMlr8DeCbO
80LORVUBBBcTBzaPxmmnfrlXC3bu3Dzw
rxODabL5H6LkFhXhov0iCYKaxTA9SKSY
MYzsHdeN4rEc6ne1SobyUoK2u2bedJUp
sUDb7jmMMO6h7QCGGbkwa7LRv1JYu8hy
5C1M33g2KLtBshvnj3V6S2MYFxbvFgbr
S8dOGQgMx9WOH90orySWSfGuBL2mpxgC
UldfKBDGVlNpqUbcVPLV68eOFqucUj06
zjSqovtw62tjDA2tSelpFpAvfXbGUyI5
dxAGfKahLvCc0GhMyac32CmQo10JGLuF
3rt8nhT9Ddda15KAPVhRXJgPcilmGipU
5mhIirR785Vve6LjBZyOqRviHmWUZl9M
kxrBCJhec2Aub2FmnrU1dCIx7f3H1ST6
rYbtzQzJJVg8Wksx9dU8wHg2X1R8zulJ
yAP2PJs1dDSFdYXe6GrQtQC7iO8oyX3L
hrBQwXe1RNG3VJXIAQAgC7JAjl2XlHyD
PKOt8NcGbxRcfL8tTxbrdxJe6zY1WS9f
FRU8DMa3f1esZLFfzixxhgJkOKvG7vXX
8XdpUGiGQTqm1q6E1BiNZmYTcFwQovHP
1vxo6QPCl4cn7JrPoKgTFUd6zZ6jrN6X
HeeJJy8N63XTT1mmruOcaXwl9gH06LLR
MyE1OSaDFTs03JE5K1CLusYomSaDFufz
l1VJ9A1olk859kfiGM7UNdjqitFHl3fa
eXOCkhjYbsXlnwQwAHsFDo9XyPV9b7oj
ZizO4wnchcm4tTOACaPUrkRtopGIjkFq
P45k7fhnHTEWYkEMfEPSAK2tp2hizA7O†
mkcvU3fDfRGULYMwqmOFCUlqPTzrSb8P
RhDBHQQCPFEvVYagp1SorSLbEygFUiAL
C2Lt7JOOpGRwdu2TrCMDNn1jE8nBCHpb
KkgHbcRejP3vgmwPVpI3jW3Pq6cddRsJ
F8Uu1bzONcullMqx3o7S1o3M4VRg7EtR
52JOaQM1ksTl9M7xo3AYi3Hto6exeWpG
nozzxBTVeanjyN8aDAOknM5gv55sydOG
t1dRq19eE7PizNhNBod8AT5pX18KFAul
Yp6NRNtFSjbjdcFWN7Q1ATwOir9wrBNw
Oa7cBsB7PQPy7Kr8jHkJorml61mv4kIP
srARDl4a3Z4Gtb39GTOv0ynNVE6T7xHS
KX83zwqAkUKsTWQKeWmfjewObf8y7upu
wuxfEVb6iJ7GghVUjN5eKTO0VsulXaOg
gm7PhGPwaWaEeOoGOlKOaltS9tLOBYGt
zWtiCjFSxBFRmU3DklC4UMFKFHCOXJgS
xc54eJI7x5BwwtOLSfdRnYdvMLagr1sp
RqC1HNwuWs5fhqEheSYlRc93EoReggVm
U2onxpoiCTT2F4SHmDMTeWb2GgWtRZhb
Adm1jCIZpVR5Evov0BZ5XyEA3QXgkpiO
niVBznqEcE3RfyWa2u7EXjpguXBMB9dE
nrtZyE7IBmlm0ZH3gG7FHJAoYxlFqczB†
x6nbOt0fqHhNHbYCgQP0Eaa7ymp6eKuq
QHauJpImWK54mPysmthbCSRI6BwQfuim

Table 20: Validation set

I4PBCtY88EMiljTpJ6ns5kIoimlIUgfl
ip25KNFw4RbDNeQTXjLI95OsnQLLe35e
vvcikOyQkXvfZdNkFYEWveZjUlQmYrSf
euzXQcpDnB1xsKeKOJHYW481SNZXiVHK
AavMakHhFeoHz9AgdWVTBsXCdiOLBNEk
KBn6Fz4XRNh5A7YRBBGZIlSVzcmV6P51
AEhNZrB5mb9K7hCBxB0xvJzUC5WxqaFN
ssRHUDqjbEgtepnEnjPpOd6WQHekD6PZ
XpSmjNEWOMad8655KQ4q3NjDHNUNHW6C
9q7wGfr4xVuiodCCrqViuS1dnZ8tthvZ
A2YTNgCrkIoMGvSkyzK5ROEFgXVRfptP
xijlZSyZXb5MEfbKLM54iOJRJkRdrBrX
OCi1DrQQgJ7GjNFNTrxscclxkaDHOY76
yWLp2YwYQlKqoSR14JUiazZzcywX8Xj8
jusHjPORbNBFg6iMvaNE4HcFMc4oJyqD
3veUj2KxK6jmJlllRj9tXbSncK6xnSOx
ZwojnSyDEhD6s8ZERY8AoDWL0cnBH7BU
GJShIRZ3lsuqeG7HhgpwlMw5v8prIuD4
ZFkXoBHziu00Toye78g1YdKm5P0bA0dY
pY4EUZQszT0kL5Et76FkEbRlLeAMT2hz
oTlKy2ISbZfN7i2jTcZ4mCqmCpK6dhDm
TLYaLIWCrBbtqLySvqxj67g5ZAfn2Zhs†
MT9Dkn3L0akMVov0lRUq9HmNsP49R1dX
rHg2gQNoevGYPUag6PAn9CANXKmx17ms
Vth5VKxswvoQEpLCis2O0xjSOf0ctjFu
vc3ShhPeF5CiTm9HiOmoakicyNWNulab
XSedsGunPd3IUuZ8RJUGmz3SUoCv0rzW
Pk9lFIEExot3gjvV83ZDTuWSj6dqk9uW
tihrah4T4i2gA8dscroj9Mu5715fXojg
M7d3C4ZQtx0RO0Wum7JMxQtZExUfNEzi†

Table 21: Test set

vHJMSYFIl1TfLweQ5DWMGN5f47ULFNxe
MT1zW5iBOh1bxF42QBpyDqotQk7NcnHw
leySrSnra9yAO3eTJIGB55nrjRS3RqIW
FLsgQZoIGHx1G3LmdD7jtICMik2EKRKN
YFQX33c8EEoapTndd2084KbUuUmtj7xF
KB84bUmLWOrKCKkISCn8QuNBhF5mgOL8
oEJ7fslCTL7s0OfIe7nYIPqo7Il4rMjI
CzwqE37s81YahjNICSXI2Tb4Fmp6bclE
Su02hndUSYGSKJmcSqroKmtDjXIJ4y60
LcsapTberZwzU7qyEr11andO59HTOVCv
0vjSLbBj2sckqQam3tZ92QLDpQNYqaa0†
mXHZZ887A9fcZgOmnxhnPVHwu5EC1jDG
RpARZ17l5osnFUcqIj2aTOsgRBMDutoA
tIgyhF8T1BOZnu7h6jb58igU5MAGgdo9
kKDzUlAprXqDz84Nrw9UP1W0jpgUKkhN

This research was funded by STZ Softwaretechnik GmbH, part of the IT-Designers Gruppe, Esslingen am Neckar, Germany. It contains all samples that have been collected during the clinical study.

Any other comments?
None.

J.2. Composition

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of instances (e.g., nodes, edges) present in the dataset? Please provide a description.
The instances in the dataset represent diagnostic polysomnographic sleep recordings, which include up to 36 raw and 23 derived data channels, alongside 81 types of annotated events for each participant, supplemented by data from various questionnaires.

How many instances are there in total (of each type, if appropriate)?
The dataset encompasses 113 diagnostic polysomnographic sleep recordings.

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not.
The dataset is a sample of diagnostic sleep recordings from adult patients undergoing regular and purely diagnostic examinations at the sleep laboratory of the Klinik für Kardiologie, Angiologie und Pneumologie of the medical clinic in Esslingen am Neckar, Germany. It contains all samples that have been collected during the clinical study. The representativeness compared to other datasets on arousal diagnostics is discussed in Appendix H.

What data does each instance consist of? "Raw" data (e.g., unprocessed text or images) or features? In either case, please provide a description.
Each instance consists of raw polysomnographic data including up to 36 raw and 23 derived data channels, 81 types of annotated events, and data from various questionnaires. Descriptions of all channels and fields that are used in this work are provided in Appendix H. A full description of the dataset is available on the PhysioNet page of the CPS dataset (Kraft et al., 2024).

Is there a label or target associated with each instance? If so, please provide a description.
Yes, each recording includes labels for sleep-related events such as arousals, apnea, hypopnea, and other sleep events as per the American Association of Sleep Medicine (AASM) guidelines (Berry et al., 2012). A full list of annotated events is available on the PhysioNet page of the CPS dataset (Kraft et al., 2024).

Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.
The Pittsburgh Sleep Quality Index (PSQI) questionnaire was only given to 62 patients. The remaining 51 patients did not receive the questionnaire, so this information is missing for those instances. All questionnaires contain missing values for some questions due to non-response. Raw data is complete for all patients, but five patients are missing derived systolic and diastolic blood pressure channels.

Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.
No explicit relationships between individual instances are made.

Are there recommended data splits (e.g., training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.
We provide recommended splits for training, validation, and test data, listed in Appendix I and on the PhysioNet page of the CPS dataset (Kraft et al., 2024).

Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.
The dataset may contain some noise typical of polysomnographic data, such as artifacts from patient movement or external interference, but efforts were made to minimize these by conducting comprehensive quality assurance in the pilot phase of the clinical study. The channels *Heart Rate*, *Light*, and *SpO2* are smoothed versions of the raw data channels *Pulse*, *Light*, and *SPO2*, respectively, processed using the DOMINO software from SOMNOmedics GmbH.

Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? If external resources are required, please describe them, as well as any restric-

tions (e.g., licenses, fees) associated with them.
The dataset is self-contained.

Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals' non-public communications)? If so, please provide a description.
Yes, the dataset contains confidential patient data protected under doctor-patient confidentiality agreements. The most sensitive data are the sleep medical diagnoses. The dataset has been anonymized to protect patient identities.

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.
No, the dataset does not contain such data.

Does the dataset identify any subpopulations (e.g., by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.
The dataset includes demographic information such as age and gender to study their impact on sleep-related arousals. It was determined from questionnaires. The age distribution in years is: Under 50: 26 patients (23.01%), 50-60: 34 patients (30.09%), 60-70: 23 patients (20.35%), over 70: 22 patients (19.45%), unknown: 8 patients (7.08%). The approximate gender distribution, based on 62 patients, is: Male: 45 patients (72.58%), Female: 17 patients (27.42%). Gender information is only available in the aggregated statistics, not for individual patients, as an anonymization measure.

Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset? If so, please describe how.
No, patient identities are anonymized to prevent identification.

Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, locations of health data about individuals or genetic data, forms of financial information, such as social security numbers, salary)? If so, please provide a description.
The most sensitive data are the sleep medical diagnoses, consisting of short extracted textual descriptions of patients' sleep-related medical conditions from doctors' letters, e.g., obstructive sleep apnea with severity indications. Statistics and a complete listing of the sleep medical diagnoses are available in the official documentation on the PhysioNet page of the CPS dataset (Kraft et al., 2024) and in the file *statistics.html* in the supplementary material (generated using the SweetViz (Bertrand, 2020) library, under MIT license).

Any other comments?
None.

J.3. Collection Process

How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie rating), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part of speech tags, model-based guesses for age or language)? If the data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.
The data was directly observable from polysomnographic recordings and supplemented by data from various questionnaires.

What mechanisms or procedures were used to collect the data (e.g., hardware apparatuses or sensors, manual human curation, software programs, software APIs)? How were these mechanisms or procedures validated?
The data collection during polysomnographic examinations involved SOMNOscreen devices and the DOMINO software, both from SOMNOmedics GmbH, Germany. The data was validated, curated, and extended with additional labels (e.g., sleep stages and arousals) by trained sleep medical scorers from NRI Medizintechnik GmbH (Germany) following guidelines from the American Academy of Sleep Medicine (AASM). The data was exported from DOMINO in EDF (raw data) and TXT (annotations) formats. An employee of Klinikum Esslingen (funded by IT-Designers Gruppe) performed the digitalization of the questionnaires and doctor's letter in YAML-format, the pseudonymization of the whole data and made the data available to the research group from IT-Designers Gruppe. The data was then further anonymized for release. The whole process was developed in collaboration with data protection officers from Klinikum Esslingen and IT-Designers Gruppe. It was approved by the ethics committee of the Landesärztekammer Baden-Württemberg,

Germany. The data quality was validated in a pilot phase.

If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?
The dataset was collected monocentrically at the sleep laboratory of the Klinik für Kardiologie, Angiologie und Pneumologie in Esslingen am Neckar, Germany, ensuring a representative sample of adult patients undergoing regular diagnostic examinations.

Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?
The data collection was conducted by clinical staff at Klinikum Esslingen. One student was hired by the clinik to perform the digitalization and pseudonymization of the data. He was paid 12.98 EUR per hour. The expenses for the student were covered by STZ Softwaretechnik GmbH.

Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances? If not, please describe the timeframe in which the data associated with the instances was created.
The data was collected during 2021-2022.

Were any ethical review processes conducted (e.g., by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.
The study protocol was approved by the ethics committee of the Landesärztekammer Baden-Württemberg, Germany, on 2020-10-21 (committee number F-2020-105, https://www.aerztekammer-bw.de/ethikkommission).
The clinical study was registered at the German Clinical Trials Register, DRKS-ID: DRKS00033641 (Wienhausen-Wilke and Kraft, 2024).

Does the dataset relate to people? If not, you may skip the remaining questions in this section.
Yes, the dataset relates to people.

Did you collect the data from individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?
The data was collected directly from individuals at the Klinikum Esslingen sleep laboratory.

Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.
Patients were informed as part of the clinical study consent process. The consent form followed a template and was approved by the ethics committee of the federal state, the Landesärztekammer Baden-Württemberg, Germany.

Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.
Yes, patients gave informed consent for the collection and use of their data. The content of the clinical study and all general conditions were explained verbally by the medical staff and in writing in the informed consent form. It was administered in a preliminary visit prior to the examination.

If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to, or otherwise reproduce, any supporting documentation.
Patients were informed of their right to revoke consent at any time.

Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.
Yes, an impact analysis was conducted with a focus on the risks of re-identification, data misuse, data breaches, and failure to achieve the study objectives. Outcomes (measures) to reduce the risks were as follows: Selection of an established platform for data sharing (PhysioNet) including usage limitations (future usage must be in line with the original study goals) and the requirement of a data use agreement and public credentialed access. Additionally, in order to anonymize the data, we removed most free text, the sensitive attributes medication and pre-existing conditions, and multiple indirect identifiers like gender and profession that were less important for achieving the study goals. For the remaining indirect identifiers (age and BMI), we selected bins that respected k-anonymity with k=3. For the medical diagnosis (the remaining sensitive attribute), we made sure to have i-diversity with i=2 among the k-

anonymous groups. Absolute timestamps in the measurement data were adjusted to start on January 1st, 1970. Our measures cover the criteria required by the safe harbor method from the Health Insurance Insurance Portability and Accountability Act (HIPAA) and go beyond them.

Any other comments?
None.

J.4. Preprocessing/cleaning/labeling

Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remaining questions in this section.
Derived data channels and some event data were automatically calculated by the DOMINO software from SOMNOmedics GmbH. Manual labeling was performed by trained sleep medical scorers from NRI Medizintechnik GmbH following guidelines from the American Academy of Sleep Medicine (AASM). Bucketing of age and BMI attributes and shifting of timestamps were performed to anonymize the data. Apart from this, all raw data were converted from EDF to WFDB format, which is the standard format for PhysioNet. In the process, all raw data was upsampled to the highest sampling rate of 256 Hz.

Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data? If so, please provide a link or other access point to the "raw" data.
Access to the raw data beyond the clinical study is limited to the medical clinic, Klinikum Esslingen.

Is the software used to preprocess/clean/label the data available? If so, please provide a link or other access point.
The software that was used for preprocessing is proprietary. The DOMINO software from SOMNOmedics GmbH can be licensed from the company.

Any other comments?
None.

J.5. Uses

Has the dataset been used for any tasks already? If so, please provide a description.
This publication entails the first use of the dataset.

Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.
No.

What (other) tasks could the dataset be used for?
The dataset may be used for research that aligns with the original study goals (Wienhausen-Wilke and Kraft, 2024). The study is aimed at investigating the utility of Machine Learning (ML) for improving the quality and efficiency of sleep-related arousal diagnostics, reducing technical demands of the data collection process, assessing the utility of a transparent clinical decision support system, studying the clinical relevance of arousals on sleep quality, and the utility of ML for medical knowledge discovery in this context.

Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might limit its usability for other tasks?
We do not foresee any limitations on the usability of the dataset for the tasks mentioned above.

Any other comments?
None.

J.6. Distribution

Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.
Yes, the dataset is publicly available for research purposes with credentialed access on PhysioNet (Kraft et al., 2024).

How will the dataset be distributed (e.g., tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?
The dataset is available on PhysioNet (Kraft et al., 2024).

When will the dataset be distributed?
The dataset is already available on PhysioNet (Kraft et al., 2024).

Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.
The dataset is distributed under the PhysioNet Credentialed Health Data License 1.5.0[2].

Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a

2. https://physionet.org/about/licenses/physionet-credentialed-health-data-license-150/

link or other access point to, or otherwise reproduce, any supporting documentation.
There are no third-party IP-based restrictions.

Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.
There are no export controls or other regulatory restrictions.

Any other comments?
None.

J.7. Maintenance

Who will be supporting/hosting/maintaining the dataset?
The dataset is hosted on the PhysioNet platform (Kraft et al., 2024). Support and maintenance will be provided by the authors of this publication.

How can the owner/curator/manager of the dataset be contacted (e.g., email address)?
Contact information for the dataset maintainers can be found in the official documentation on the PhysioNet page of the CPS dataset (Kraft et al., 2024) under the *Corresponding Author* section.

Is there an erratum? If so, please provide a link or other access point.
Currently, there is no erratum. If the need for an erratum arises, the dataset can be updated on PhysioNet with semantic versioning.

Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? If so, please describe how often, and how updates will be communicated to dataset consumers (e.g., mailing list, GitHub)?
Updates will be deployed as necessary to correct any errors. Communication will be done via PhysioNet.

If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were the individuals in question told that their data would be retained for a fixed period of time, or were they told that they could have their data deleted)? If so, please describe these limits and explain how these limits will be enforced.
Anonymized data will be retained indefinitely on PhysioNet. Access to pre-anonymized data is restricted to the IT-Designers Gruppe for four years after the end of the data collection, as communicated to study participants. Klinikum Esslingen will retain the original data in accordance with legal requirements. Deletion requests only affect pre-anonymized data. Since the patient IDs in the anonymization process were created randomly and not linked to any patient information, deletion of the anonymized data of a specific patient is technically not possible.

Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to dataset consumers.
Older versions will be maintained on PhysioNet to ensure continuity and reproducibility of research.

If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description.
We are not aware of any mechanism on PhysioNet for users to contribute directly to the dataset. However, we are open to collaboration and will consider any requests for extensions or contributions.

Any other comments?
None.

Appendix K. Utilized compute and environmental impact

Experiments following the *DeepSleep* approach ran on single Nvidia GeForce RTX2080TI GPUs with 11 GB of memory. They are part of a workstation containing four GPUs. The workstation is equipped with an Intel Core i9-10900X CPU with 10 cores and 20 threads and 128 GB of DDR4 RAM. Experiments using models implemented in the *sktime* library (Löning et al., 2019) ran on the CPU of the same workstation.

Table 22 gives an overview of the average runtime for single experiments conducted for this publication, where we also indicate how often experiments were repeated to obtain confidence intervals.

In total, we have 391.17 GPU hours and 20.26 hours without GPU usage. From this, we conduct estimations of $kgCO_2eq$ for the GPU hours using the MachineLearning Impact calculator[3] presented in Lacoste et al. (2019). We use a factor of 0.4880 $kgCO_2eq$ per kWh for the electricity mix in Germany.

Total emissions are estimated to be 47.72 $kgCO_2eq$. Due to preliminary exploratory experiments and repetitions of experiments due to changes in requirements or fixes, we estimate the total emmission to be thrice as high, i.e. about 150 $kgCO_2eq$.

3. https://mlco2.github.io/impact#compute

Table 22: Runtime and count of experiments conducted for this publication.

	Experiment	Environment	Repetitions	Average runtime [h]
Main experiments on the CPS dataset	D4	GPU	5	10.8
	D3	GPU	5	6.5
	D2	GPU	5	0.7
	D1	GPU	5	0.7
	IndividualBOSS	CPU	5	0.18
	SupervisedTimeSeriesForest	CPU	5	1.4
	TimeSeriesForestClassifier	CPU	5	0.7
	SignatureClassifier	CPU	5	1.1
	SummaryClassifier	CPU	5	0.2
	Catch22Classifier	CPU	5	0.4
	RandomStratified	CPU	5	0.03
	RandomUniform	CPU	5	0.03
	Constant 1	CPU	1	0.03
	Constant 0	CPU	1	0.03
2018 PhysioNet Challenge	Target, IOD	GPU	5	11.6
	Target, POD	GPU	1	23,75
	Target, FED	GPU	5	6.8
	Non-target, IOD	GPU	5	10,6
	Non-target, POD	GPU	1	33,2
	Non-target, FED	GPU	5	11,9
MISC	Ablation study	GPU	30	0.65
	HP Tuning	GPU	44	0.38

D Assessing the Real-World Utility of Explainable AI for Arousal Diagnostics: An Application-Grounded User Study

The full text of the following publication is included in this appendix:

Stefan Kraft, Andreas Theissler, Vera Wienhausen-Wilke, Gjergji Kasneci, and Hendrik Lensch. Assessing the real-world utility of explainable ai for arousal diagnostics: An application-grounded user study. *arXiv preprint arXiv:2510.21389*, 2025

Assessing the Real-World Utility of Explainable AI for Arousal Diagnostics: An Application-Grounded User Study

Stefan Kraft*
IT-Designers Gruppe
Esslingen am Neckar, GER
stefan.kraft@it-designers.de

Andreas Theissler
University of Giessen
Giessen, GER
https://orcid.org/0000-0003-0746-0424

Vera Wienhausen-Wilke
Klinikum Esslingen, Klinik für Kardiologie, Pneumologie und Angilologie
Esslingen am Neckar, GER
v.wienhausen-wilke@klinikum-esslingen.de

Gjergji Kasneci
Technical University of Munich
Munich, GER
gjergji.kasneci@tum.de

Hendrik Lensch
University of Tübingen
Tübingen, GER
hendrik.lensch@uni-tuebingen.de

Abstract

Artificial intelligence (AI) systems increasingly match or surpass human experts in biomedical signal interpretation. However, their effective integration into clinical practice requires more than high predictive accuracy. Clinicians must discern *when* and *why* to trust algorithmic recommendations. This work presents an application-grounded user study with eight professional sleep medicine practitioners, who score nocturnal arousal events in polysomnographic data under three conditions: (i) manual scoring, (ii) black-box (BB) AI assistance, and (iii) transparent white-box (WB) AI assistance. Assistance is provided either from the *start* of scoring or as a post-hoc quality-control (*QC*) review. We systematically evaluate how the type and timing of assistance influence event-level and clinically most relevant count-based performance, time requirements, and user experience. When evaluated against the clinical standard used to train the AI, both AI and human-AI teams significantly outperform unaided experts, with collaboration also reducing inter-rater variability. Notably, transparent AI assistance applied as a targeted QC step yields median event-level performance improvements of approximately 30% over black-box assistance, and QC timing further enhances count-based outcomes. While WB and QC approaches increase the time required for scoring, start-time assistance is faster and preferred by most participants. Participants overwhelmingly favor transparency, with seven out of eight expressing willingness to adopt the system with minor or no modifications. In summary, strategically timed transparent AI assistance effectively balances accuracy and clinical efficiency,

*Also: University of Tübingen, GER

providing a promising pathway toward trustworthy AI integration and user acceptance in clinical workflows.

1 Introduction

Artificial Intelligence (AI) has demonstrated impressive diagnostic capabilities in medicine, often achieving or surpassing expert-level performance in fields such as radiology, pathology, dermatology, ophthalmology, and cardiology (Topol, 2019). However, successful deployment in clinical settings depends not only on high predictive accuracy but also on meaningful integration into medical workflows (Fawzy et al., 2023).

In sleep medicine, one key diagnostic task is the analysis polysomnographic (PSG) data, where the scoring of arousal events – brief electroencephalographic (EEG)-based disruptions of sleep – is essential for diagnosing sleep disorders such as obstructive sleep apnea (OSA) or periodic limb movement syndrome (Berry et al., 2012; Franklin and Lindberg, 2015). OSA alone affects approximately 20% of the population and is associated with severe health consequences (Franklin and Lindberg, 2015; Wetter et al., 2012).

Arousal scoring, traditionally performed by trained experts, is a time-consuming process with high inter-rater variability, particularly across different institutions (Chylinski et al., 2020; Ehrlich et al., 2024). This variability in scoring quality, combined with high demand for PSG examinations, leading to long waiting times, creates an urgent need for reliable clinical decision support systems (CDSS) that can augment human expertise while maintaining clinical standards. In response, deep learning models like *DeepSleep* have emerged, achieving state-of-the-art performance on public PSG datasets and showing promise for clinical application (Li and Guan, 2021; Ehrlich et al., 2024).

Despite these advances, real-world adoption of AI-based CDSS is hindered by the lack of interpretability, understandability, and transparency, which undermine accountability and reduce trust in predictive outcomes (Fawzy et al., 2023).

Explainability and the evaluation gap. Explainable AI (XAI) offers the potential to enhance decision-makers' validation capabilities, improve outcome quality, foster trust, and facilitate accountability in collaborative decision-making (Mersha et al., 2024).

However, the effectiveness of explanations in clinical settings remains poorly understood, with conflicting evidence from existing studies. Recent XAI research reveals that explanations don't always improve human-AI collaboration performance. For instance, Bansal et al. (2021) found that explanations are beneficial only when AI systems outperform humans working independently. Additionally, studies by Poursabzi-Sangdeh et al. (2021) and Panigutti et al. (2022) indicate that explanations can hinder error detection due to automation bias. Furthermore, Schmidt and Biessmann (2020) demonstrated that explanations of incorrect predictions can increase algorithmic bias among risk-averse users.

The evaluation of the utility of AI explanations is therefore a crucial aspect of explainable AI (XAI) research, and there is ongoing debate about the best approach to measure the effectiveness of these explanations (Amarasinghe et al., 2023). One of the most widely accepted methods is the use of *application-grounded* evaluations, which involve real users performing real tasks in real-world environments (Amarasinghe et al., 2023; Doshi-Velez and Kim, 2017). They are often considered the gold standard for evaluating AI explanations (Gunning et al., 2021) because they offer the most authentic assessment of how explanations affect real-world outcomes. By involving domain experts in actual tasks, these studies provide insights into the practical utility of explanations in professional environments, where the stakes are often high.

Yet, only about 20% of XAI evaluations apply user studies (Rong et al., 2022; Nauta et al., 2023) at all and only about 5% conduct application-grounded user studies (Rong et al., 2022; Nauta et al., 2023). The rest of the evaluations are *functionally-* or *human-grounded* evaluations with proxy tasks or lay participants or are even just evaluated with anecdotal evidence (Nauta et al., 2023). This scarcity of real-world, task-specific evaluations leaves a significant gap in understanding how well AI

explanations actually perform in practice, especially in high-stakes environments where expert users are essential.

We assert that compelling evidence for the utility of AI explanations demands *application-grounded* studies that integrate objective task metrics with subjective user feedback within the intended environment. This methodology should adhere to guidelines derived from best practices in human-AI interaction, as discussed for example by Rong et al. (2022). The significance of these best practices is underscored in recent research by Amarasinghe et al. (2024), who illustrates how selecting evaluation metrics that align with business objectives and acknowledge real-world decision-making options, such as deferring a decision to a senior expert, can fundamentally alter the conclusions drawn about the utility of explanations.

Contributions of this work. This study addresses this gap in the evaluation of the utility of AI explanations for the high-stakes task of arousal scoring and makes three specific contributions:

1. We incorporate *DeepSleep*, a state-of-the-art arousal detection tool, into a web-based Decision Support System, specifically designed for the task of arousal scoring. This system can function as a black box or provide comprehensive explanations presented with varying levels of detail.

2. We design and conduct an application-grounded user study involving eight professional sleep scorers. For each participant, the study spans approximately twelve hours of physiological timeseries data for arousal event scoring and includes detailed questionnaire data.

3. We examine how *transparency* (white- vs. black box) and *timing* (assistance from the start vs. post-hoc quality control) influence (i) diagnostic performance against two complementary ground truths, (ii) time efficiency, and (iii) user-centred factors such as trust, comfort, and perceived plausibility.

Research questions. Our investigation is structured around three research questions:

RQ1: Does AI assistance improve human arousal scoring performance and efficiency relative to unaided experts?

RQ2: Does transparent (*white box*) assistance outperform opaque (*black box*) assistance in terms of objective performance and subjective acceptance?

RQ3: When is assistance most beneficial? During the initial scoring pass or as a post-hoc quality control step?

1.1 Related Work.

Evolution of Automated Arousal Detection. Early approaches to automated arousal detection were based on traditional machine learning techniques using hand-crafted features (Zan and Yildiz, 2023). However, these methods were quickly outperformed by deep learning models utilizing end-to-end architectures (Li and Guan, 2021). A major catalyst in the field was the 2018 PhysioNet Challenge on arousal detection (Ghassemi et al., 2018; Goldberger et al., 2000), where the winning solution introduced *DeepSleep*, a fully convolutional neural network architecture (Li and Guan, 2021), setting new standards for segmentation accuracy and throughput. Subsequent research has built on these foundations, with continued innovation in deep learning architectures driving further improvements in detection performance (Zan and Yildiz, 2023; Kuo et al., 2023; Badiei et al., 2023).

Despite these technical advances, the evaluation of arousal detection models has historically prioritized methodological benchmarks over clinical applicability. Only recently has attention turned to assessing these models in real-world settings (Ehrlich et al., 2024; Kraft et al., 2025). Notably, previous evaluation protocols often disregarded prevalent clinical practices, such as annotating only arousal onset points rather than complete events, and were not guided by clinically relevant considerations (Kraft et al., 2025). This resulted in fragmented evaluation practices, relying on technically motivated window-based or pointwise schemes, while overlooking clinically important factors such as accurate event counts and the selection of meaningful decision thresholds (Kraft et al., 2025).

Most Similar Prior Work: Ehrlich et al. 2024. Among existing studies, the work by Ehrlich et al. (2024) represents the most thorough and clinically-relevant assessment of automated arousal detection to date. Their research bridges the divide between technical performance and clinical utility by comprehensively evaluating a U-Net-based detector derived from the *DeepSleep* architecture (Li and Guan, 2021). Their evaluations span both a large clinical dataset and several public datasets, with evaluation measures such as the arousal index (ArI) error employed to reflect real-world relevance.

This study overcame several limitations of prior research, specifically restricted and non-diverse patient samples and the use of clinically irrelevant evaluation metrics. They addressed the gap in understanding model generalizability across patient groups by leveraging a large, heterogeneous, real-world dataset collected from routine sleep laboratory practice (Ehrlich et al., 2024). Findings from robust cross-dataset testing revealed significant performance shifts, underscoring the necessity of multi-center data and rigorous external validation, while the strategy of dataset mixing was noted to be promising for enhancing generalization in the future. By selecting the ArI error as their primary evaluation measure, the authors strengthened the methodological alignment of their study with clinical practice.

While their results reaffirm the considerable promise of automated arousal detection, they also emphasize persistent challenges: low inter-rater reliability, limited diversity among patient populations, and pronounced variation in cross-dataset performance.

The relevance of this prior work to our study is considerable, given that we likewise adopt the *DeepSleep* architecture, determine the accuray of total arousal count estimation (which is similar to them utilizing the ArI error measure), and maintain a focus on clinical effectiveness, including similar observations regarding scorer reliability and inter-standard performance divergence. Nonetheless, our approach extends this foundation by conducting a fully application-grounded user study, an element not addressed by Ehrlich et al. (2024). Furthermore, our research explicitly examines dimensions of AI transparency, such as post-hoc attribution and the visualization of confidence scores and decision threshold. Moreover, in contrast to their study, our previous work (Kraft et al., 2025) made our collected dataset publicly available. Finally, our evaluation uniquely incorporates direct feedback from professional scorers, thereby enabling a holistic assessment of clinical utility and user acceptance. This allows us to systematically quantify the impact of transparency and the timing of AI assistance on accuracy, bias, efficiency, and user acceptance, factors directly tied to clinical utility.

The Unclear Value of Explainability and the Need for Application-Grounded Evaluations. The prevailing argument suggests that enhancing the explainability of AI-driven decision support systems can foster user trust and optimize human-AI collaboration in medical contexts (Yu et al., 2025). Yet, empirical evidence emphasizes that the benefits of explainability cannot be assumed to be universal or automatic (Rosenbacke et al., 2024; Gaube et al., 2023). For example, a recent systematic review by Rosenbacke et al. (2024) found that clinicians' trust in AI systems often increases with clear and clinically relevant explanations, particularly when such explanations resonate with medical intuition. Conversely, ambiguous or overly technical explanations may have the opposite effect, leading to diminished trust.

Beyond trust, explainability can also directly influence diagnostic performance and collaborative efficacy. In an AI-assisted radiography study, Prinster et al. (2024) demonstrated that the format of the explanation plays a critical role: radiologists who received local, feature-based explanations improved in both accuracy and receptivity to the AI's recommendations compared to those provided global, prototype-based explanations. Notably, this influence on reliance and trust occurred largely below participants' conscious awareness, indicating that explanatory modality subtly shapes clinician behavior. Meanwhile, research by Gaube et al. (2023) showed that clinicians with less experience derived the greatest benefit, both objectively and subjectively, from AI explanations, with experienced experts not receiving any significant benefit from explanations.

On the other hand, explainability can inadvertently encourage over-reliance. Buçinca et al. (2021) observed that users often deferred to AI suggestions, even erroneous ones, unless the user interface incorporated "cognitive forcing" mechanisms to prompt deliberate review. The complexity

of evaluating explanations is further heightened by phenomena such as the perception-behavior gap (Gaube et al., 2021), where clinicians may report skepticism towards AI-generated advice but nonetheless rely on it in practice, revealing a divergence between stated preferences and actual behaviors.

Collectively, these studies underscore the importance of engaging healthcare professionals in controlled, context-rich experimental designs that mirror real-world workflow. Doing so not only surfaces usability and interaction issues potentially overlooked in technical evaluations, but also more accurately documents how explanations affect clinically relevant outcomes. The literature increasingly advocates for application-grounded evaluation strategies that consider workflow compatibility, cognitive burden, and concrete decision impacts, rather than presuming that additional explanation invariably produces greater benefit (Rosenbacke et al., 2024; Gaube et al., 2021).

In this context, our work aims to address these challenges by systematically investigating how AI transparency, contrasting transparent (white-box) with opaque (black-box) assistance, affect trust, efficiency, and diagnostic accuracy among professional sleep scorers in arousal scoring.

1.2 Paper outline.

This paper is organized as follows: Section 2 elaborates on the methodologies employed in this study. Section 3 describes the experimental design, including participant recruitment, data and model selection, and the study protocol. Section 4 presents both quantitative and qualitative findings for each research question. Section 5 interprets the results comprehensively, discusses limitations, and concludes the study.

2 Methodologies

This section presents the core methodologies applied in this study. Section 2.1 offers an overview of the machine learning model architecture, training, and inference strategies for arousal detection. Section 2.2 introduces the explanation approaches, encompassing both local (Section 2.2.1) and global (Section 2.2.2) post-hoc feature attribution techniques. Section 2.3 details the construction of a consensus ground truth from multiple annotators, including annotation clustering (Section 2.3.1) and the Expectation-Maximization method for joint estimation of true labels and annotator quality (Section 2.3.2). Section 2.4 describes the performance evaluation framework, addressing both event-based (Section 2.4.1) and count-based (Section 2.4.2) metrics. Section 2.5 explains the pairwise inference procedures using the CPS ground truth and consensus labels. Finally, Section 2.6 outlines the statistical analysis framework, including experimental design, modeling strategies, repeated-measures ANOVA, permutation testing, effect size estimation, and simple effects analysis. Table 17 (Appendix A) provides a table of notation.

2.1 Machine Learning Model

Detecting arousal events in sleep is challenging due to their brief, subtle nature. Classical machine learning (ML) models often require extensive feature engineering and may not generalize well (Zan and Yildiz, 2023). To address these issues, we employ a ML model designed for robust, end-to-end arousal onset detection. Specifically, we use the top-performing model, referred to as *D4* in our prior research (Kraft et al., 2025).

This section provides a concise overview of the model's architecture, training, and inference processes. For comprehensive technical details, including hyperparameter configurations, readers are directed to our previously published work (Kraft et al., 2025).

Our model is a modified version of the DeepSleep architecture, specifically adapted for detecting arousal onsets in sleep data (Li and Guan, 2021). As refined by Fonod (2022), it employs a streamlined U-Net architecture, reducing the depth from 11 to 5 layers to decrease computational demands while maintaining performance. A weighted binary cross-entropy (BCE) loss function addresses data imbalance, with arousal events defined as 10-second intervals around the onset, in accordance with

clinical scoring guidelines (Berry et al., 2012). Detailed descriptions of the model architecture and training process are available in earlier publications (Fonod, 2022; Kraft et al., 2025).

During inference, the model produces confidence scores for each time step, representing the likelihood of an arousal event initiation. To mitigate noise and reduce false positives, the output is smoothed using a 3-second averaging filter, following the methodology outlined in Kraft et al. (2025).

Binary predictions are derived from these confidence scores using the Approximate Localization and Precise Event Count (ALPEC) framework (Kraft et al., 2025). ALPEC facilitates post-processing and performance evaluation for arousal detection by ensuring precise event localization and counting. The ALPEC variant applied here determines the decision threshold based on the F2-score from the training set, prioritizing the minimization of missed events (false negatives). This focus is crucial in clinical contexts, where recalling arousal events is more critical than precision to prevent missing significant occurrences. The F2-score is highlighted as the primary metric for performance evaluation, as it aligns with the operational objectives of clinical decision support systems, where minimizing false negatives is essential, allowing clinicians to address false positives, thereby optimizing resource use while maintaining high efficacy (Deo, 2015). Following thresholding, predicted intervals within 10 seconds of each other are merged, intervals exceeding 60 seconds are discarded, and a 15-second temporal buffer is applied on both sides of the ground-truth annotations to ensure clinically relevant alignment.

For an in-depth understanding of the ALPEC framework's rationale and its alignment with clinical standards, refer to Kraft et al. (2025).

2.2 Explanation Methods

The DeepLift method (Shrikumar et al., 2017) is employed to generate both local and global post-hoc feature attributions. It is a gradient-based interpretability technique recognized for its effectiveness in neural time series classifiers (Šimić et al., 2025). Also, it provides notable advantages in computational efficiency and memory usage over many alternative methods. DeepLift operates by comparing the activation of a neuron to a reference activation, with the resulting difference used to assess the significance of input features in the model's decision-making process. The choice of the reference activation will be discussed in the context of the study design in Section 3.4.1. The following subsections briefly describe the methodologies applied in this study. The underlying rationale and illustrative examples are presented in Sections 3.4.1 and 3.4.2, respectively.

2.2.1 Local Explanations

To provide local explanations for individual arousal events while minimizing computational demands and complexity, we focus on calculating attributions solely for the time point identified by the model as the most probable event onset. We present visualizations of the attributions for the top 10 channels deemed most significant to the model's decision-making process. Channel importance is determined by aggregating attributions across all time points. We display attributions for a 60-second window centered symmetrically around the predicted event onset. Furthermore, we exclusively present positive attributions that exceed a predefined threshold, thereby simplifying the explanation.

2.2.2 Global Explanation

In this study, a global explanation of the model's decision-making process is derived by aggregating feature attributions across all events from subjects included in the test set that are selected in the user study. The attributions are summed over all time points and all events for each channel. Subsequently, channels are ranked based on their importance for the model's decision. The 20 channels with the highest importance are depicted in a bar chart for visualization.

2.3 Consensus Ground Truth Derivation

To adequately assess the performance of both human annotators and the AI model in identifying arousal events, we want to establish a consensus ground truth that encapsulates the collective expertise of multiple annotators. Given the temporal nature of arousal events and the potential misalignment of annotations from different experts, we adopt a two-step approach: (1) clustering annotations based on their temporal proximity, and (2) employing an Expectation-Maximization (EM) algorithm to estimate both the consensus ground truth and the performance of individual annotators. Both steps are described in the following subsections.

2.3.1 Clustering of Annotations

Each annotator provides a series of arousal event start times, denoted as $\{t_{k,i}\}$, where k represents the annotator and i signifies their respective events. To group annotations that likely correspond to the same underlying event, we employ temporal clustering.

Given the irregular temporal distribution of arousal events, fixed temporal windows are unsuitable, and overlapping windows would unnecessarily complicate the analysis. Therefore, we organize the annotations into clusters $\{C_j\}$ of similar events, with each cluster C_j comprising annotations from one or more annotators that are temporally close to each other.

To mitigate potential bias from selecting a single clustering algorithm, we utilize two distinct methods: DBSCAN (Ester et al., 1996), a density-based clustering algorithm, and agglomerative clustering (Sokal and Michener, 1958), a hierarchical clustering algorithm. The event start times $t_{k,i}$ constitute the one-dimensional feature space for clustering.

DBSCAN

The DBSCAN algorithm effectively identifies clusters of densely packed annotations while designating isolated points as noise. The optimal value for the ε parameter is determined through the k-distance method in conjunction with the Kneedle algorithm (Satopaa et al., 2011). To ensure robust parameterization, we evaluate multiple sensitivity values S for the Kneedle algorithm, which modulate the algorithm's responsiveness to knee point detection. Lower sensitivity values make the algorithm less responsive, typically identifying knee points later in the curve (resulting in larger ε), whereas higher sensitivity values detect knee points earlier (yielding smaller ε). Subsequent post-processing ensures that each cluster includes no more than one annotation per annotator and that all clusters meet a minimum annotator count.

Agglomerative Clustering

Agglomerative clustering is employed to iteratively merge annotations based on their temporal proximity. This process utilizes Ward's linkage (Ward Jr, 1963) as the distance metric. Post-processing steps are implemented to eliminate duplicate annotations from the same annotator within a cluster and to exclude clusters that lack a sufficient number of unique annotators.

2.3.2 Expectation-Maximization for Simultaneous Truth and Annotator-Quality Estimation

Following the clustering phase, we utilize the Expectation-Maximization (EM) technique to simultaneously estimate the true occurrences of events and evaluate annotator performance. This process involves adapting the well-established Dawid-Skene EM algorithm (Dawid and Skene, 1979) to our specific context of clustered, time-stamped event annotations. Our aim is to determine the probability that each cluster represents a true event while also assessing the sensitivity and specificity of each annotator. This is accomplished through a two-class latent-variable model, wherein the latent variable signifies the presence of an event within a cluster. Each cluster is modeled to potentially contain either zero or one latent true event.

Initialization

Sensitivities (s_k) and specificities (p_k) for each annotator k are initialized to 0.99, reflecting an initial assumption of high performance while avoiding computational issues associated with starting at 1.0. The initial probability that a cluster C_j represents a true event is set to $P_j = 0.5$.

EM Iterations

E-Step (Expectation): For each cluster C_j, we compute the probability P_j that it represents a true event, given the current estimates of annotator sensitivities s_k and specificities p_k. For each annotator, we define $A_{k,j} = 1$ if they contributed to cluster C_j, and 0 otherwise. The log-likelihood of the observed annotations in C_j if an event is truly present is given by

$$\log(L_j) = \sum_{k=1}^{K} \left[A_{k,j} \log(s_k) + (1 - A_{k,j}) \log(1 - s_k) \right]. \tag{1}$$

Conversely, the log-likelihood if no event is present is given by

$$\log(M_j) = \sum_{k=1}^{K} \left[A_{k,j} \log(1 - p_k) + (1 - A_{k,j}) \log(p_k) \right]. \tag{2}$$

The posterior probability for an event in cluster C_j is then

$$P_j = \frac{\exp(\log(D_{1j}))}{\exp(\log(D_{1j})) + \exp(\log(D_{2j}))}, \tag{3}$$

where the log-posterior probabilities for an event being present and absent are given by $\log(D_{1j}) = \log(L_j) + \log(P_j^{\text{prev}})$ and $\log(D_{2j}) = \log(M_j) + \log(1 - P_j^{\text{prev}})$.

Here, P_j^{prev} and $1 - P_j^{\text{prev}}$ are the prior probabilities of an event being present / absent in cluster C_j from the previous iteration.

M-Step (Maximization): We update the sensitivities s_k and specificities p_k of the annotators based on the current estimates of cluster probabilities as

$$s_k = \frac{\sum_j P_j A_{k,j}}{\sum_j P_j}, \tag{4}$$

and

$$p_k = \frac{\sum_j (1 - P_j)(1 - A_{k,j})}{\sum_j (1 - P_j)}, \tag{5}$$

These updates reflect the expected true positive and true negative rates, weighted by the current cluster probabilities.

Convergence Criteria: The algorithm iterates until the maximum change in any cluster probability falls below a set tolerance (e.g., 1×10^{-7}), or a maximum number of iterations is reached.

Probability Bounds: After each update, all probabilities and performance parameters are clipped to the $[0, 1]$ interval to ensure validity.

Consensus Event Selection: After convergence, clusters with $P_j \geq \tau$ are considered consensus events, where typically $\tau = 0.5$ The consensus event time for each cluster is computed as the mean of the annotated times within the cluster.

Implementation Notes: In our implementation, all calculations are performed in log-space to ensure numerical stability, effectively mitigating potential underflow and overflow issues associated with extremely small or large numbers.

Furthermore, the noise cluster is deliberately excluded from all Expectation-Maximization (EM) updates to preserve the integrity of the results.

The method is implemented in Python, utilizing the robust functionalities of libraries such as NumPy and scikit-learn.

To further enhance the robustness of our calculations, we introduce small constants, such as 1×10^{-10}, during logarithmic computations or divisions. This precautionary measure prevents undefined operations, such as the logarithm of zero or division by zero, thereby ensuring the reliability of the computational process.

Critical Considerations

The accuracy of the consensus is fundamentally dependent on the efficacy of the clustering process. Poorly constructed clusters can substantially compromise the precision of the consensus outcomes.

Moreover, the algorithm presupposes the independence of annotators, a common assumption in analogous methodologies. Nonetheless, this assumption may not consistently reflect real-world conditions, thereby potentially affecting the reliability of the results.

2.4 Performance Evaluation

Performance is assessed in two primary ways: the evaluation of individual event scoring, detailed in Section 2.4.1, and the evaluation based on the total event count, elaborated in Section 2.4.2.

2.4.1 Event-Based Performance Evaluation

The evaluation of an expert's performance, which may be a human annotator, the AI model, or a human-AI collaborative team, is conducted by assessing the alignment between expert annotations and a specified ground-truth.

For human experts, the onset point of each annotated event serves as the reference point for comparison. In the case of the AI model, the argmax values of the model's confidence score output for each event are utilized as reference points. For human-AI teams, reference points are determined by either the onset points of events marked by humans or the argmax value of an AI event that has been accepted by the human annotator.

We proceed depending upon the type of ground-truth employed:

Consensus Ground-Truth The center point of a cluster is calculated as the mean value of the annotations constituting the cluster, which is then used as the reference point for comparison.

For the evaluation of human solo performance, calculations could be based on the existing clustering structure. However, for novel annotations not previously included in the clustering procedure, a method must be established to associate them with either a consensus cluster or a noise cluster. A pragmatic approach involves utilizing the distance of an annotation to the centroids of consensus clusters to determine the nearest cluster. This is done by defining a distance threshold, derived empirically from clustering statistics, which determines whether an annotation should be assigned to a consensus cluster or classified as noise. For consistency, human solo performance will also be evaluated using this methodology.

CPS Ground Truth An externally determined standard, in the form of ground-truth event annotations, may also be employed for comparison. In this context, the ground truth from the CPS dataset, that was used to train and evaluate the AI model, is utilized. The onset of an event annotation serves as the reference point for comparison to expert annotations.

The evaluation involves two binary sequences: one for expert predictions and another for ground-truth annotations, where events are denoted by a single 1 at the reference point indicating an event start, and 0 elsewhere.

Matches between expert and ground-truth annotations are determined as follows: a predicted 1 is counted as a true positive (TP) if it occurs within a specified time window, defined by the distance threshold on either side of a ground-truth 1. Predicted 1s that fall outside this window are counted as false positives (FP). Conversely, any ground-truth 1 that does not have a corresponding predicted 1 within the threshold window is counted as a false negative (FN).

Unlike the AI model, where the primary optimization and performance evaluation metric was the F2 score (refer to Section 2.1), the F1 score, which represents the harmonic mean of precision and recall, is now employed for a more balanced evaluation. Additionally, the F2 score, precision, recall, as well as TP, FP, and FN counts are reported for a comprehensive performance analysis.

Benefit Ratio. The *benefit ratio* is a normalized alignment score utilized to evaluate the performance of a human-AI team in comparison to a fixed ground truth. The benefit ratio is applicable in scenarios where the AI outperforms the unaided human, that is, when $F1^{\mathrm{AI}} > F1^{\mathrm{HU}}$. The benefit ratio is calculated by normalizing the team's performance relative to the performance of an unaided human and the AI's performance.

The formula for the benefit ratio $\mathcal{B}$ is defined as:

$$\mathcal{B} = \frac{F1^{\mathrm{HU+AI}} - F1^{\mathrm{HU}}}{F1^{\mathrm{AI}} - F1^{\mathrm{HU}}}, \tag{6}$$

where $F1^{\mathrm{HU+AI}}$ is the F1 score of the human-AI team, $F1^{\mathrm{HU}}$ is the F1 score of the unaided human (solo performance). A value of $\mathcal{B} > 1$ would indicate that the human-AI team has performed better than the AI alone, while a value of $\mathcal{B} < 0$ would indicate that the human-AI team has performed worse than the human alone. In the case, where $F1^{\mathrm{HU}} < F1^{\mathrm{HU+AI}} < F1^{\mathrm{AI}}$, the benefit ratio is bounded between 0 and 1. In this case, an elevated benefit ratio signifies that the human-AI team has effectively harnessed the AI's capabilities to enhance performance beyond the level attainable by the human alone, in comparison to the AI's maximum potential contribution.

In scenarios where the AI model is trained on the CPS ground truth, the benefit ratio provides insight into the extent to which the human participant has adopted the AI's alignment towards this standard.

Relative F1 Scores. In cases where we don't have access to the human solo performance, we can still determine the alignment of the human-AI team to the AI's performance. Comparing the raw F1 scores of a human-AI team across different experimental human-AI interaction regimes is methodologically flawed when the baseline performance of the AI varies between these regimes. For example, if the AI demonstrates sub-optimal performance in one condition, the human-AI team may appear less effective in absolute terms. This apparent weakness is not necessarily due to inadequate human performance but rather because the AI's enhanced performance reduces the potential for further improvement.

To address this issue, we use the relative F1 scores $\mathcal{R}$, defined as the ratio of the F1 score of the human-AI team to that of the AI alone.

$$\mathcal{R} = \frac{F1_{\mathrm{HU+AI}}}{F1_{\mathrm{AI}}} \tag{7}$$

This normalization quantifies the human-AI team's capacity to leverage the AI's potential relative to its stand-alone performance, thereby controlling for regime-specific differences in baseline difficulty or suitability. This interpretation of alignment is most appropriate in scenarios where the AI outperforms the human-AI team, that is, when $F1^{\mathrm{AI}} > F1^{\mathrm{HU+AI}}$. Nevertheless, it should be interpreted with caution: a high alignment score neither guarantees greater absolute accuracy nor implies proximity to a clinical ground truth. Instead, it indicates that human corrections remain close to the AI's initial proposal, which may reflect trust, compliance, or alignment, but not necessarily improvement. In this sense, $\mathcal{R}$ is best understood as a proxy for *calibrated cooperation* rather than performance per se. Its validity rests on the critical assumption that the AI baseline approximates a well-curated, consensus-driven ground truth.

2.4.2 Count-Based Performance Measures

This section introduces count-based measures per recording

Notation. For each recording let C_{GT} be the ground-truth arousal count. For each source $x \in \{\mathrm{AI}, \mathrm{HU+AI}, \mathrm{HU}\}$, let C_x denote its corresponding arousal count. Here, AI refers to the AI-only output, HU+AI to the human-AI team output, and HU to the human-solo output, which is available exclusively in QC regimes. The primitive absolute count deviation used throughout is defined as

$$D_{x \to \mathrm{GT}} = \left| C_x - C_{\mathrm{GT}} \right|. \tag{8}$$

Count Accuracy (Bounded APE). As a bounded, interpretable score we use the inverse absolute percentage error (APE) transform

$$A_{x,\mathrm{GT}} = \frac{1}{1 + \mathrm{APE}_{x,\mathrm{GT}}}, \qquad \mathrm{APE}_{x,\mathrm{GT}} = \frac{D_{x\to\mathrm{GT}}}{\max\{C_{\mathrm{GT}}, \epsilon\}}, \quad \epsilon > 0. \tag{9}$$

Here, we set $\epsilon = 10^{-8}$ to ensure numerical stability. The accuracy measure $A_{x,\mathrm{GT}}$ ranges within $(0, 1]$, attaining a value of 1 under perfect agreement and decreasing monotonically as the absolute percentage error increases.

AI-Baseline Improvement Ratio. Our primary improvement measure compares how much the team closes the GT gap relative to the AI on a ratio scale:

$$R_{\mathrm{GT}} = \frac{D_{\mathrm{AI}\to\mathrm{GT}} + \delta}{D_{\mathrm{HU+AI}\to\mathrm{GT}} + \delta}, \qquad \delta > 0. \tag{10}$$

We use $\delta = 10^{-6}$ to ensure numerical stability and analyze its log form

$$y_{\mathrm{RGT}} = \log R_{\mathrm{GT}} = \log\big(D_{\mathrm{AI}\to\mathrm{GT}} + \delta\big) - \log\big(D_{\mathrm{HU+AI}\to\mathrm{GT}} + \delta\big), \tag{11}$$

so that $y_{\mathrm{RGT}} > 0$ ($R_{\mathrm{GT}} > 1$) indicates the team is closer to ground truth than the AI by a multiplicative factor R_{GT}, $y_{\mathrm{RGT}} = 0$ parity, and $y_{\mathrm{RGT}} < 0$ a worse team outcome. For reporting we back-transform point estimates and confidence intervals via $\exp(y_{\mathrm{RGT}})$.

Percentage Error. To assess systematic count bias toward over- or under-counting, we report the team's percentage error:

$$\mathrm{PE} = \frac{C_{\mathrm{HU+AI}} - C_{\mathrm{GT}}}{\max\{C_{\mathrm{GT}}, \epsilon\}}, \tag{12}$$

We set $\epsilon = 10^{-8}$ to ensure numerical stability. A value of $\mathrm{PE} > 0$ indicates over-counting, while $\mathrm{PE} < 0$ reflects under-counting relative to the ground truth. In contrast to R_{GT}, which quantifies only the *magnitude* of improvement, the percentage error characterizes both the *magnitude* and *direction* of the residual count error.

QC team–solo comparison (applicable only in QC regimes). In QC regimes, where a human-solo baseline is available, we directly compare $A_{\mathrm{HU+AI,GT}}$ and $A_{\mathrm{HU,GT}}$ for the same patient recordings. The paired difference, defined as $\Delta A_{\mathrm{GT}} = A_{\mathrm{HU+AI,GT}} - A_{\mathrm{HU,GT}}$, is computed to quantify the extent of improvement achieved by the team relative to the solo human performance.

Hypotheses used in inference. All tests are performed at the participant level and then aggregated per atomic regime (Start/QC × WB/BB): (i) for y_{RGT} we use the directional hypothesis $H_1 : \mu_{y_{\mathrm{RGT}}} > 0$ (team improves over AI). (ii) For bias we use the two-sided hypothesis $H_1 : \mu_B \neq 0$.

2.5 Pairwise Inference under the CPS Ground Truth (Primary) and Consensus (Sensitivity)

Scope. All confirmatory, pairwise tests use the *CPS ground truth* as the authoritative annotation source. Analyses based on the human *consensus* labels are treated as sensitivity checks and are *not* part of any multiplicity-controlled family. Because the CPS ground truth is prespecified as the sole confirmatory annotation source, each of the three questions of the second primary objective *PO2* (performance comparison of human-AI team vs. human solo, human-AI team vs. AI, and human solo vs. AI, see Section 3.6.1) forms a family of size one, so there is no multiplicity adjustment.

Outcome scales. We conduct the same inferential procedure on two evaluation scales: (i) *event-level* performance, where the per-participant outcome is the F1-score, and (ii) *count-level* performance, where the outcome is count accuracy. Let $M(\cdot)$ denote a generic per-participant performance metric (either F1-score or accuracy).

2.5.1 Human-AI Team vs. Human Solo (paired)

For each participant $i = 1, \ldots, n$ we form the paired difference

$$d_i \;=\; M^{\mathrm{CPS}}_{\mathrm{HU+AI},i} - M^{\mathrm{CPS}}_{\mathrm{HU},i}. \tag{13}$$

We test $H_0\colon\ \mu_d = 0$ (where μ_d denotes the population mean difference) using (a) the paired t-test and (b) an exact, two-sided sign-flip permutation test on the *mean* difference $\bar{d}$. With $n \in \{8, 9\}$ across both standards, we enumerate all 2^n sign configurations exactly. We report $\bar{d}$ with a nonparametric percentile bootstrap 95% CI based on 10,000 resamples of $\{d_i\}_{i=1}^{n}$. The identical workflow is applied to the consensus labels for sensitivity only.

2.5.2 Human-AI Team and Human Solo vs. AI (one-sample vs. scalar)

Let $a^{\mathrm{CPS}}_{\mathrm{AI}}$ be the AI's performance under CPS ground truth (a single scalar). For each participant i we define two comparisons:

$$d_i^{\mathrm{HU+AI}} \;=\; M^{\mathrm{CPS}}_{\mathrm{HU+AI},i} - a^{\mathrm{CPS}}_{\mathrm{AI}}, \tag{14}$$

$$d_i^{\mathrm{HU}} \;=\; M^{\mathrm{CPS}}_{\mathrm{HU},i} - a^{\mathrm{CPS}}_{\mathrm{AI}}. \tag{15}$$

For each comparison, we test $H_0\colon\ \mu_d = 0$ (where μ_d denotes the population mean difference) using a one-sample t-test and an exact two-sided sign-flip permutation test on $\bar{d}$, and report $\bar{d}$ CIs and d_z with bootstrap CIs. Consensus-based sensitivity analyses contrast $M^{\mathrm{cons}}_{\mathrm{HU+AI},i}$ and $M^{\mathrm{cons}}_{\mathrm{HU},i}$ to $a^{\mathrm{cons}}_{\mathrm{AI}}$ in the same manner.

2.6 Statistical Inference for Team Performance over all Regimes under the CPS Ground Truth

2.6.1 Experimental Design and Scope of Inference

All inferential analyses employ a 2×2 within-participant factorial design, with the factors *Timing* (levels: Start and QC) and *AI transparency* (levels: WB and BB; hereafter referred to as *AI*). The study includes eight participants who provide complete data across all four experimental conditions. Inferential results are confined to these four atomic regimes (Start/QC × WB/BB). Composite regimes, which combine data from multiple phases or conditions (indicated with a "+" symbol, e.g. "Start,BB+WB"), are used solely for descriptive purposes and are excluded from hypothesis testing to preserve statistical independence and maintain a clearly defined family of tests.

2.6.2 Statistical Modeling and Analysis Scale

All statistical models are fitted on the logarithmic scale to stabilize variance and facilitate additive modeling of multiplicative effects. Outcome variables are log-transformed with a small offset ($\varepsilon = 10^{-8}$) to address zero values. All point estimates and 95% confidence intervals are reported as back-transformed ratios using $\exp(\cdot)$. Sphericity is not a concern, as each within-subject term in the 2×2 design has one degree of freedom.

2.6.3 Repeated-Measures ANOVA and Planned Contrasts

Repeated-measures ANOVA (RM-ANOVA) is conducted, with subject as the repeated-measures factor and *Timing* and *AI* as within-participant factors. To test the study hypotheses, standard orthogonal contrasts are defined for the 2×2 design and applied to participant-wise cell means: (i) the main effect of *AI* (WB vs BB), (ii) the main effect of *Timing* (Start vs QC), and (iii) the interaction effect.

2.6.4 Permutation Testing and Multiple Comparisons

For each contrast, participant-wise contrast scores are computed and their means are tested using within-subject contrast t-tests. Exact sign-flip permutation p-values are provided, with Holm correction applied within contrast families. Specifically, Holm correction is used for the two main effects, while the interaction effect is evaluated using unadjusted permutation p-values. This approach ensures robustness in the presence of small sample sizes and effectively controls the familywise error rate.

2.6.5 Effect Sizes and Confidence Intervals

Effect ratios (back-transformed contrast means) and the within-subject effect size d_z are reported. Confidence intervals for both measures are obtained via nonparametric bootstrap (10,000 resamples of participant-level contrast scores, using the percentile method). Effect sizes are presented as both partial eta-squared (η_p^2) from ANOVA and Cohen's d_z from contrasts: η_p^2 quantifies the proportion of variance explained by the factor, while d_z represents the standardized mean difference for the contrast.

2.6.6 Simple Effects Analysis

When significant interactions are identified, simple effects are examined using paired t-tests with sign-flip permutation tests and Holm correction within the family of simple effects. This procedure clarifies the nature of the interaction by comparing conditions within each level of the other factor.

3 User Study Design

This section provides a comprehensive overview of the user study design. We begin by outlining the general design considerations that guided the development of the study in Section 3.1. Next, we describe the overall structure and objectives of the arousal scoring task in Section 3.2. The design and implementation of the Decision Support System (DSS) are detailed in Section 3.3, followed by an explanation of the different modes of explainability available to participants in Section 3.4. Section 3.5 details the dataset and sample selection process, while Section 3.6 provides an in-depth description of the study protocol including objectives, recruitment and participants and the conrete procedure of the study. The overarching evaluation approach and its underlying rationale are then outlined in Section 3.7.

3.1 General Design Considerations

As highlighted in Section 1, application-grounded user studies are vital for assessing the practical utility of AI explanations. Consequently, Amarasinghe et al. (2023) emphasize that evaluations should be conducted within the context of the specific task the AI system is intended to support. To achieve this, we utilized contemporary patient data sourced from a sleep clinic, enlisted the expertise of professional sleep scorers, developed a web-based application enabling these scorers to execute the arousal scoring task, and employed evaluation measures pertinent to the clinical environment.

3.2 Task Overview

Participants in our study are tasked with scoring arousal events using polysomnography data.

This task is an essential element of sleep diagnostics, based on the American Academy of Sleep Medicine guidelines (Berry et al., 2012). Professional sleep scorers, who are trained in visually analyzing polysomnography (PSG) data, execute this task by identifying the onset of arousal events. In clinical settings, typically only the onset of an arousal event is marked, as its duration is generally deemed clinically irrelevant. For simplicity, we will refer to these as arousal events, although the emphasis is on identifying the onset.

The AI-powered Decision Support System (DSS) identifies regions where the onset of an arousal event is most probable. If a scorer agrees that an arousal begins within the suggested region, they may accept the event without further modification. Alternatively, they may create an event independently. If the AI-suggested region is ambiguous and the scorer wishes to accept the AI event, they must adjust the region to ensure it encompasses the correct onset of the arousal. In the absence of an AI-determined region near the scorer's perceived arousal onset, they must manually create a new arousal event.

Upon completion of the scoring process, all manually created arousal events, along with accepted AI-suggested events, are included in the final scoring. There is no obligation to explicitly reject AI-suggested events that the scorer disagrees with.

Furthermore, for each identified arousal event, whether manually created or accepted from AI suggestions, participants are encouraged to provide a confidence assessment. This assessment is a subjective evaluation of the scorer's confidence in the accuracy of their identification and can be selected from the following five levels: "very uncertain", "uncertain", "confident", "very confident", and "fully confident", with "confident" as the default. The definitions of these levels, which are covered during the initial training session of the user study, are as follows:

Very uncertain signifies a slight inclination towards believing the event is an arousal but with considerable doubt.

Confident denotes a standard level of confidence in the event being an arousal.

Fully confident indicates the scorer has absolute certainty that the event is an arousal, with an expectation that this assessment is highly unlikely to be incorrect.

Uncertain and Very confident serve as intermediate levels, offering gradations between the other confidence levels.

In the context of this user study, the arousal scoring task is simplified by not requiring causal differentiation between various arousal event types, and arousal scoring is concluded at 1 AM for each patient. According to the guidelines of the American Academy of Sleep Medicine, all arousals, regardless of their cause, must be visible in the electroencephalogram channel. Therefore, identifying a generic arousal class remains a realistic task routinely performed by professional sleep scorers.

3.3 Decision Support System

The Decision Support System (DSS) is a web-based platform that facilitates the arousal scoring task through an intuitive user interface. It is designed to enable users to conduct the arousal scoring task with or without AI assistance. Figure 1 shows the main interface of the DSS.

The development of the DSS was conducted in collaboration with a seasoned sleep medicine expert, who also possesses extensive experience as a trainer of medical scorers.

The design of the Decision Support System (DSS) was guided by several key principles to ensure both functionality and usability. Central to the system is the integration of an efficient machine learning (ML) model for the detection of arousals during sleep, as described in Section 2.1. This model was trained on patient data from the Comprehensive Polysomnography Dataset (Kraft et al., 2024), ensuring clinical relevance and robustness. In addition to its predictive capabilities, the DSS is equipped to provide comprehensive explanations for the ML model's predictions, a feature that is further elaborated in Section 3.4.

A primary objective in the system's development was to achieve efficient time behavior, enabling rapid responses to user inputs and timely data processing. The platform was designed to be intuitive and easy to learn, with an attractive user interface that minimizes the learning curve and enhances the overall user experience. To further support user efficiency, the interface was engineered to minimize the number of required interactions, presenting all relevant information within a single view and offering keyboard shortcuts for common actions as well as multiple navigation options, such as quickly moving between intervals or events.

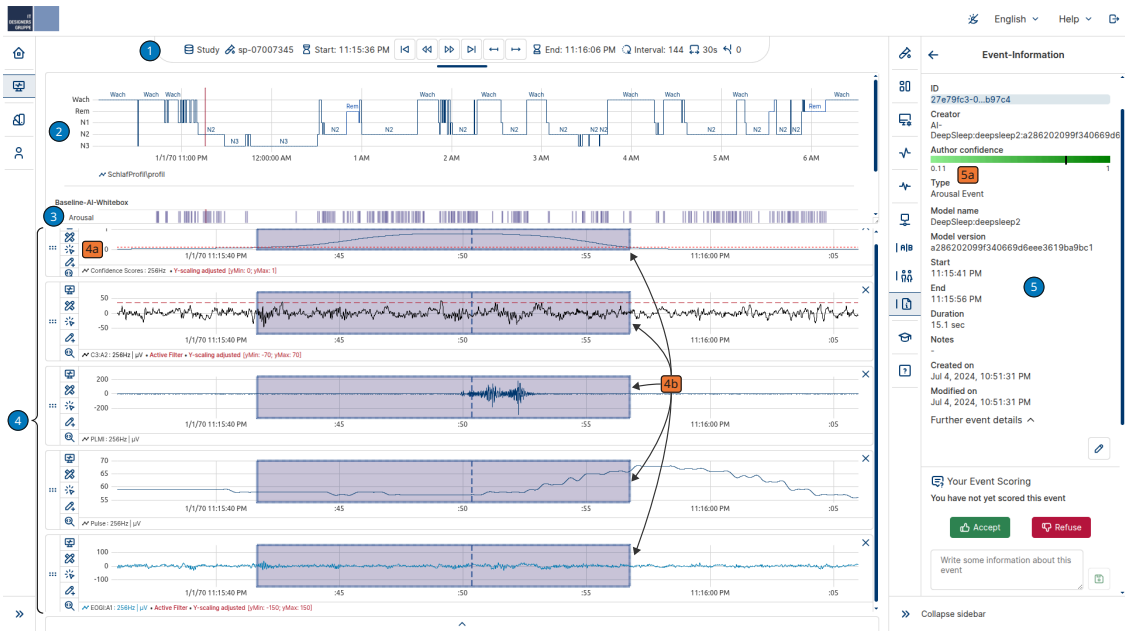


Figure 1: **Main interface of the Decision Support System** for the arousal scoring task. The interface is organized into several key areas, each marked with blue circles: **1)** Control bar providing access to basic information, and navigation tools; **2)** Overview graph displaying the hypnogram (sleep stages) for the entire recording, with the currently selected interval highlighted by an orange vertical bar; **3)** Timeline summarizing AI-annotated arousal event positions across the recording; **4)** Visualization of selected polysomnography data channels for the current interval, including overlays for AI-suggested arousal regions; **5)** Event information panel presenting details of the currently selected arousal event, including options to accept or reject the event; Additional transparency elements, highlighted with orange ovals, include: **4a)** Confidence score channel, visualizing the AI model's confidence for arousal onset at each time point; **4b)** Shaded regions on each channel, indicating intervals where an arousal onset is likely, with the most probable onset marked by a vertical dashed line (corresponding to the maximum confidence score); **5a)** Bar indicator showing the maximum confidence score for the selected event, visualized with a green gradient.

Customization was another important consideration, allowing users to tailor the interface to their preferences by adjusting channel configurations, color schemes, and other settings. These configurations can be saved as profiles and reloaded as needed, supporting both flexibility and consistency in user workflows. The DSS also incorporates specific features such as auxiliary lines within channels and filtering options for time series data, including high-pass and low-pass filters. Finally, the system provides comprehensive event management capabilities, enabling users to either manually annotate events or accept or refuse AI-suggested events, thereby supporting a seamless integration of human expertise and AI assistance.

3.4 Modes of explainability

The DSS incorporates multiple complementary modes of explainability to support AI-assisted arousal scoring, each designed to enhance user understanding and trust in the system's predictions.

1. **Local explanations** are provided via post-hoc feature attributions for the most probable onset of arousal events, enabling users to interpret the model's decision at the individual event level. For a detailed description and illustration, see Section 3.4.1 and Figure 2.
2. A **global explanation** is presented as a bar chart summarizing the overall importance of each data channel in the model's decision-making process, offering users insight into which physiological signals most strongly influence predictions across the dataset. Further details are provided in Section 3.4.2 and Figure 3.
3. The DSS enables users to select a data channel that visualizes the AI model's **confidence scores** for each time point, highlighting regions with a high likelihood of arousal onset. This functionality is depicted in Figure 1, area 4a).
4. The **decision threshold** used by the model is explicitly visualized within the confidence score channel, allowing users to see how predicted scores relate to the threshold for arousal detection (see Figure 1, area 4a)).
5. The **most probable onset** of an arousal event is clearly marked within the predicted interval, providing a precise temporal reference for users (see Figure 1, area 4b)).
6. The DSS also offers a concise visualization of the **confidence value** associated with the most likely onset of a predicted arousal event, supporting rapid assessment of prediction certainty (see Figure 1, area 5a)).

3.4.1 Local Explanations

Stage of Explanations: Post-hoc. We have chosen to employ post-hoc explanations that offer feature attributions for predictions. This targeted approach aligns well with the application-grounded nature of our user study, as investigating multiple forms of explanations or methods for generating post-hoc feature attributions could detract from the authenticity of a real-world setting. Furthermore, providing a singular type of explanation simplifies the process for participants, enabling us to concentrate on other critical elements, such as the timing of AI assistance. Post-hoc explanations are particularly pertinent to our study, given their prevalent use in the healthcare sector (Gupta and Seeja, 2024), thereby enhancing the relevance and applicability of our findings.

Method of Explanations: DeepLIFT. In this study, we utilize DeepLIFT for model explainability (see Section 2.2). A critical aspect of DeepLIFT is the choice of the reference activation. In other modalities of data processing, missing features are typically represented as black pixels in images or as zero/mean values in tabular datasets. However, in the context of time series data, defining a missing feature is more complex (Rezaei and Liu, 2024). To address this challenge, we employ a random Gaussian noise baseline with a mean of zero and a standard deviation of one. While this approach may introduce artifacts in the explanations for certain channels due to the different characteristics of the large number of channels in the PSG data, it remains consistent with the

objectives of our study. Our aim is not to identify new biomarkers for arousal detection. Rather, if any explanations are unfaithful, it provides an opportunity to assess the participants' ability to recognize such inconsistencies. Still, a comparison between DeepLIFT and GradientSHAP is provided in Appendix B.

Balancing Completeness and Interpretability. Following the reasoning of Gilpin et al. (2018), we recognize the necessity of balancing completeness and interpretability in explanations. While overly complex explanations can impede interpretability, offering overly simplified explanations without transparency is considered unethical. To address this, we implement three threshold levels for feature attributions. A lower threshold results in more complex yet comprehensive explanations. This information is communicated transparently to the participants.

Illustration of Local Explanations. Figure 2 illustrates the local explanation of the arousal event shown in Figure 1. During the training phase, participants are given the following info text to explain the local explanation:

> **The present graphic explains the AI model's prediction for an arousal.** The explanation allows for an understanding of the importance of individual data points and signals in predicting an arousal in the time series. It consists of several elements, described below:
>
> - **Signals:** The graphic shows the ten channels that have the highest overall relevance for the prediction, in descending order from top to bottom.
> - **Data points with relevance score above threshold X:** Blue points are placed along the signals. They mark the data points relevant to the prediction, i.e., whose relevance score is above a certain threshold (Threshold X).
> - **Threshold:** Depending on the choice of the explanation filter, more or fewer data points are marked as relevant. Choosing brief explanations/medium detail/detailed explanations leads to different thresholds. For example, a low threshold means that many data points are marked as relevant, allowing for more detailed analysis.
> - **Boundaries of the interval where an arousal is most likely to start:** Gray dashed lines indicate the interval where the AI model predicts an arousal is most likely to begin.
> - **Relevance score:** To the right of the signals is a colored scale indicating the relevance score of the data points. Dark green indicates high relevance, while lighter shades of green indicate lower relevance.
> - **Time:** The x-axis represents time, with time 0 marking the suggested arousal onset, which is the most likely time for the arousal to begin. The marked relevant data points relate to this time point.

3.4.2 Global Explanation

The global explanation of the AI model, as described in Section 2.2.2, is presented to participants as a bar chart (see Figure 3). Participants are informed that this visualization is generated by aggregating the channel relevance of all local explanations for all arousal events in the dataset, using the medium threshold setting (see Section 3.4.1).

3.5 Choice of Dataset and Samples

The dataset employed in this study is the Comprehensive Polysomnography Dataset (CPS), comprising 113 full-night sleep studies conducted during clinical operations at Klinikum Esslingen, Germany, between 2021 and 2022. This dataset encompasses up to 36 raw data channels and 23 derived channels, capturing a broad spectrum of sleep-related physiological signals. Additionally, it includes

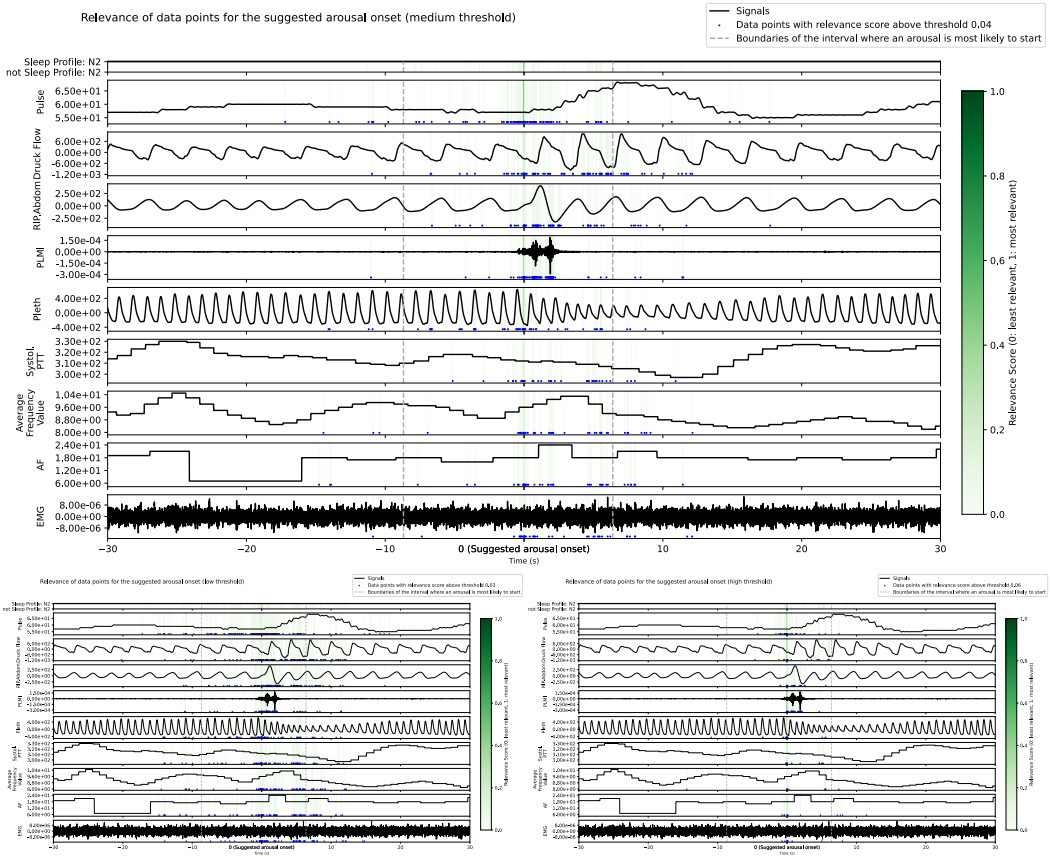


Figure 2: **Local explanations of the AI model's arousal prediction at three levels of detail** for the arousal depicted in Figure 1. Each panel visualizes the ten most relevant channels for a predicted arousal onset, with the x-axis representing time relative to the suggested arousal onset (vertical line at $t = 0$). Blue dots indicate data points with relevance scores above the current threshold, and the color bar to the right encodes the magnitude of feature relevance (dark green: high relevance, light green: lower relevance). Gray dashed lines mark the interval boundaries where an arousal is most likely to start. The top panel shows the medium detail level, while the bottom left and bottom right panels display the low and high detail levels, respectively, corresponding to different thresholds for feature attribution. Legends clarify the graphical elements.

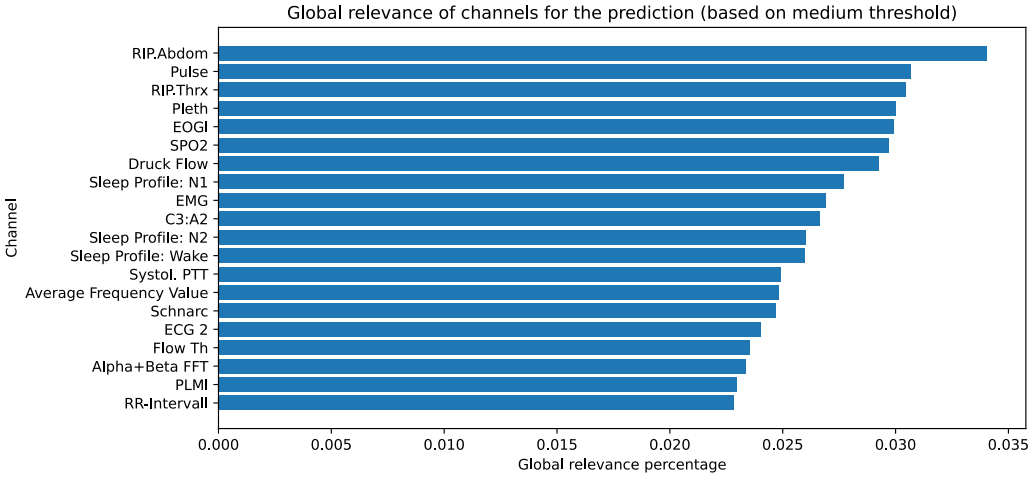


Figure 3: **Global explanation for the AI model's decision making process.** The bar chart displays the global relevance percentage of each channel for the AI model's arousal prediction, aggregated across the dataset. Channels are ranked in descending order of their contribution, with *RIP Abdomen*, *Pulse*, and *RIP Thorax* showing the highest relevance. This visualization helps users understand which physiological signals most strongly influence the model's decisions at a global level.

annotations for 81 distinct event types, such as arousals induced by respiratory disturbances, limb movements, and spontaneous arousals. For the purposes of this study, all arousal event types were consolidated into a single category to facilitate binary classification. This dataset, along with comprehensive documentation, is publicly accessible on PhysioNet (Kraft et al., 2024; Goldberger et al., 2000).

Utilizing a recently acquired real-world dataset that closely mirrors the conditions encountered by participants in their professional practice enhances the study's relevance to their expertise and the clinical environment.

In terms of preprocessing, the raw data was subjected to noise reduction using a third-order Butterworth bandpass filter, followed by normalization. The derived channels were resampled to a uniform rate of 256 Hz and normalized to a range between 0 and 1. Padding was applied to ensure uniformity across recordings. Further technical details regarding data preprocessing, hyperparameter configurations, and channel selection are available in previously published work (Kraft et al., 2025).

Sample Selection. A sample refers to a complete overnight recording of polysomnographic (PSG) data from a patient. For this study, one sample is designated for the training phase, while four samples are used for the primary segment of the user study. Table 1 presents the distribution of arousal events and associated performance metrics within the test set of the CPS dataset, ranked according to the F2-score.

The selection of samples for the user study was guided by two primary criteria: (1) The overall performance of the samples should be sufficiently high to facilitate effective learning from the AI assistance by participants. (2) The performance characteristics of the AI model across the samples should be as consistent as possible to minimize bias in the results.

A significant performance disparity in the F2-score of the AI was observed between the top-performing sample (*S1*) and the subsequent samples. Consequently, *S1* was chosen for the initial training phase of the user study. The AI model demonstrated relatively similar F2-score performance

Table 1: **Performance Metrics and Statistics of the AI Model for Patient Records in the CPS Dataset Test-Fold.** Refer to Appendix C for the mapping to actual sample IDs. Samples are ranked by their F2-score. Detailed phase descriptions are provided in Section 3.6.3.

ID	F2	F1	Precision	Recall	TP	FP	FN	#Events	Usage
S1	0.91	0.85	0.77	0.95	349	107	19	475	Phase 2 (Training)
S2	0.85	0.73	0.60	0.95	198	133	11	342	Phase 4 (QC,WB)
S3	0.80	0.77	0.73	0.82	192	71	41	304	Phase 5 (Start,BB)
S4	0.79	0.72	0.63	0.85	245	146	45	436	Phase 3 (QC,BB)
S5	0.79	0.63	0.47	0.95	184	212	9	405	-
S6	0.79	0.64	0.48	0.94	107	114	7	228	-
S7	0.74	0.72	0.68	0.75	173	81	57	311	Phase 6 (Start,WB)
S8	0.73	0.53	0.37	0.97	74	128	2	204	-
S9	0.71	0.52	0.36	0.95	173	314	9	496	-
S10	0.70	0.67	0.63	0.72	216	125	86	427	-
S11	0.68	0.71	0.77	0.66	98	30	51	179	-
S12	0.67	0.66	0.65	0.68	83	45	40	168	-
S13	0.66	0.55	0.43	0.76	119	159	38	316	-
S14	0.62	0.62	0.62	0.62	58	35	36	129	-

on the next three samples, warranting their inclusion in the study. However, two samples (*S5* and *S6*) were excluded due to the AI model's lower precision compared to the already selected samples. Thus, the next best-performing sample (*S7*) was selected for the study. The remaining samples were not included in the study.

3.5.1 Questionnaires

Participants were required to complete two detailed questionnaires designed to assess the usability, user experience, and perceived effectiveness of the AI-powered Decision Support System (DSS). These evaluations were conducted under different AI interaction regimes, specifically contrasting black box and white box modes, as well as varying the timing of AI assistance (either from the onset or as a quality control measure) within the context of the arousal scoring task. The surveys examined both the manual and the AI-assisted scoring process. The questionnaires included Likert scale items alongside open-ended questions to facilitate qualitative feedback.

The initial questionnaire was administered following the training session to collect demographic data, document any issues encountered during training, and evaluate the overall usability and user experience of the DSS. This assessment employed the standardized *AttrakDiff* questionnaire, as developed by Hassenzahl et al. (2008), which evaluates the system's pragmatic and hedonic qualities, as well as its attractiveness.

The second questionnaire was distributed at the study's conclusion. It concentrated on the perceived utility of AI assistance, participants' trust in AI predictions, their comfort level when interacting with AI, and the ease of validating AI predictions. Furthermore, it examined the plausibility of AI explanations, user preferences for upfront explanations versus AI-based quality control, and specific system functionalities, such as preferences regarding the level of detail in AI transparency and insights into potential improvements.

Collectively, these surveys offered a comprehensive perspective on user interactions with the DSS, their confidence in the system, and their views on its transparency and functionality.

3.6 Study Protocol

3.6.1 Main Objectives

As highlighted by Mohseni et al. (2021) and Rong et al. (2022), it is crucial to explicitly define design goals and evaluation criteria. In this study, we split our objectives into primary and secondary categories. The primary objectives are quantified through objective performance metrics, while the secondary objectives are evaluated via participant questionnaires. Each objective is assigned a unique identifier, such as **PO1** for the first primary objective and **SO1** for the first secondary objective. Collectively, these objectives aim to address the research questions outlined in Section 1.

Primary Objective. The primary aim of this study is to evaluate the performance of human-AI collaboration in arousal diagnostics. This involves a comparative analysis of transparent AI assistance, black box AI assistance, and manual scoring as a baseline. The specific objectives are to:

PO1 Evaluate the level of agreement among human study participants.

PO2 Assess the performance of human solo, AI solo, and human-AI collaboration using two benchmarks: (1) the consensus ground truth among study participants and (2) the CPS dataset ground truth.

PO3 Investigate the impact of different timings (AI support from the start versus AI support as quality control) and transparency modes (black box versus white box AI assistance) on the alignment of human-AI teams with the CPS dataset standard.

PO4 Analyze the time efficiency in the arousal scoring task across varying timing and transparency modes.

Secondary Objectives. For transparent AI assistance, we also want to assess:

SO1 The degree of trust participants place in AI predictions.

SO2 The perceived comfort level of participants when interacting with AI assistance.

SO3 The ease with which participants can validate AI predictions.

SO4 The plausibility of the explanations provided, including participants' preferred modes of explanation (refer to Section 3.4).

SO5 The overall satisfaction of participants when utilizing AI assistance.

SO6 The most beneficial level of detail in explanations for participants: This involves determining whether participants favor explanations highlighting only the most salient features of the model's decision or more comprehensive explanations.

SO7 The extent to which participants gain insights into the AI's general decision-making process.

SO8 The appropriateness of the AI model's decision threshold, including whether it should be adjusted to a lower, higher, or user-configurable level.

SO9 The temporal accuracy of AI predictions: This involves assessing whether the predicted arousal start intervals' proximity to the true arousal starts, as well as the intervals' duration, are sufficient for participants to evaluate the prediction's accuracy.

3.6.2 Recruitment and Participants

In studies requiring domain-specific expertise, Rong et al. (2022) highlight that evaluations by laypersons may not adequately reflect the practical effectiveness of explanations. Although lay participants are more accessible, expert participants possess the requisite skills to assess the accuracy and reliability of AI predictions, which is particularly crucial in safety-critical domains. However, recruiting experts poses challenges, often resulting in a smaller participant pool, thereby limiting the generalizability of findings, as noted by Rong et al. (2022).

During the initial recruitment phase, we encountered notable challenges which are worth mentioning. A significant number of experts declined participation, primarily due to concerns about potential job displacement by AI and the risk of disclosing expert knowledge to the public.

Despite these obstacles, we successfully recruited eight sleep scorers for the user study. The majority are employed by reputable companies such as NRI Medizintechnik GmbH and Löwenstein Medical, which offer sleep diagnostic services to hospitals and sleep laboratories. Others are independent sleep scorers. Each participant received a compensation of 300 EUR.

Figure 4 provides an overview of the study cohort's demographic composition, highlighting several notable characteristics.

The age distribution is relatively balanced, with most participants falling in the 30–40 and 41–50 age brackets, and a smaller representation from the 51–60 and 18–29 groups. Gender distribution is skewed towards female participants, reflecting the composition of the field. Regarding education, the majority of participants hold either a vocational qualification or a university degree, with a smaller subset possessing advanced academic degrees.

Most participants are employed in sleep laboratories, while a minority work in other medical contexts. Professional experience in sleep diagnostics is substantial: all participants have at least 4 years of experience, and more than half have more than 10 years. Arousal scoring experience closely mirrors this, with most participants reporting between 4 and 10 years, and several exceeding 10 years.

Experience with AI is mixed: while some participants have prior exposure to AI-based systems, others have little or no experience. Motivations for participation are diverse, with the most common being the improvement of sleep diagnostics, followed by interest in AI and technology. Additional motivations include professional development, a desire to look beyond one's immediate field, and participation due to a lack of interest from others.

This diversity in background and motivation enriches the study, providing a comprehensive view of user needs and expectations in the context of AI-assisted sleep diagnostics.

3.6.3 Procedure

The user study was conducted remotely, with participants typically working from their home offices. They accessed the Decision Support System (DSS) via a remote desktop connection to a server running the DSS. The study was facilitated through a screen-sharing session, enabling the moderator to observe and communicate with participants. The moderator intervened solely in instances of technical difficulties, protocol deviations, or to address questions about usage and configuration of the DSS.

The study adhered to a structured protocol comprising multiple phases. Participants were provided with written instructions, and new polysomnographic records were utilized in each phase to mitigate learning effects. The study employs a within-subject design, where the same test subjects are employed for all participants across the phases, allowing for a direct comparison of the different AI interaction regimes and timings per participant.

Figure 5 provides an overview of the study phases, with a comprehensive description presented subsequently.

Phases 1 and 2, which included the introduction and training, were consistently conducted at the outset. To minimize potential fatigue and learning effects, a counter-balanced design was implemented, randomizing the sequence of Phases 3 through 6 for each participant. Furthermore, to keep the study's time requirements manageable, participants scored subjects only from the beginning of the measurement until 01:00 AM during Phases 3 to 6, corresponding to approximately three hours

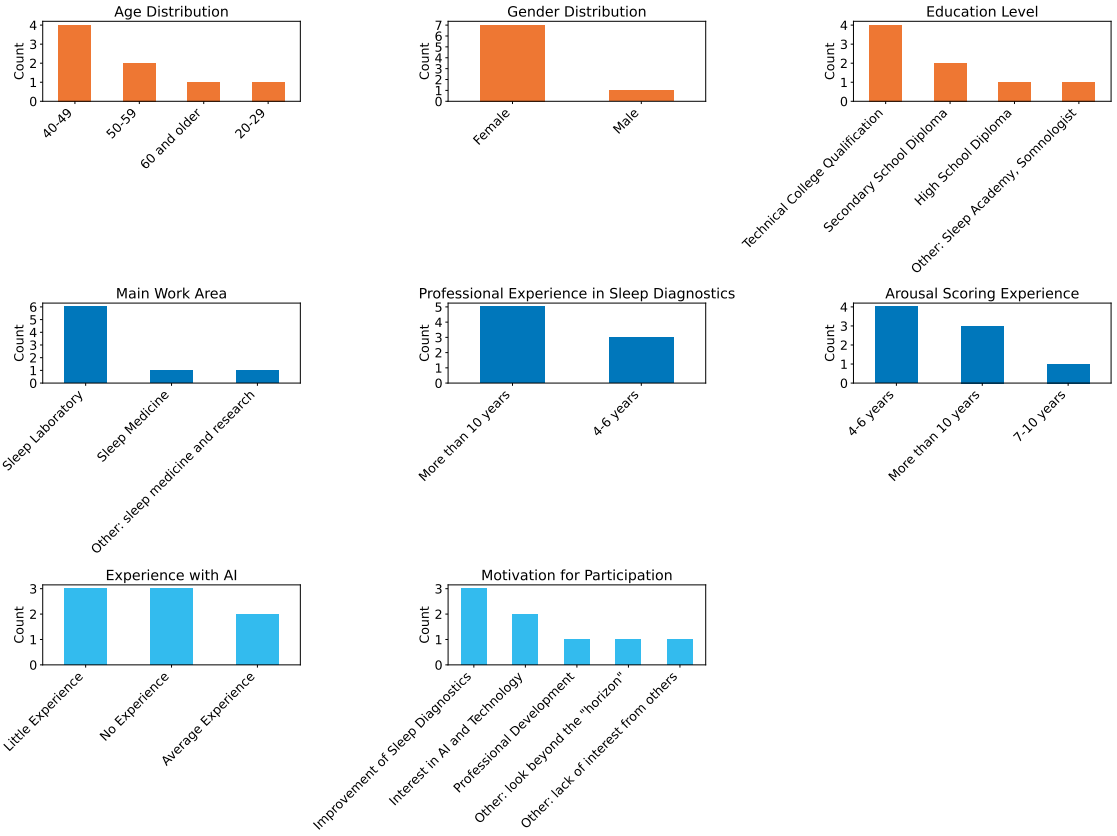


Figure 4: Detailed demographic profile of study participants, illustrating basic demographic information (orange), professional experience in sleep diagnostics (blue), and both AI experience and motivation for participation (light blue).

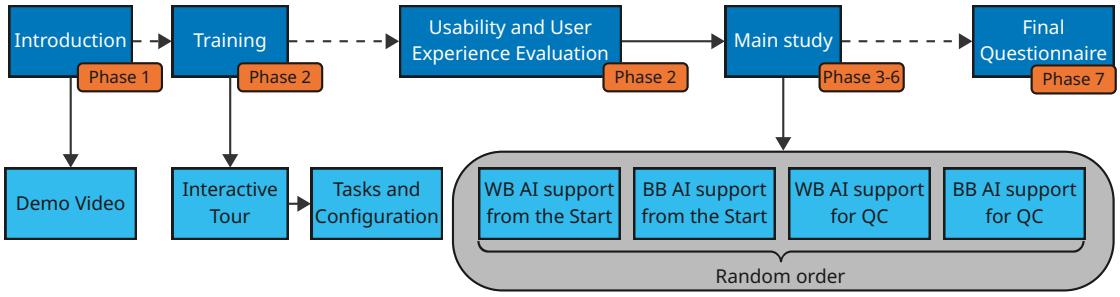


Figure 5: **Overview of the Study Phases.** The phases are annotated in orange. Dashed lines indicate that the next phase is only started after all steps of the previous phase are completed. Abbreviations: *QC* denotes Quality Control, *WB* refers to White Box, and *BB* means Black Box.

of sleep time per patient. This approach aligns closely with the application-grounded objective, as participants are not abruptly introduced to an intermediate section of a scoring. Additionally, arousals predominantly occur in the first third of the night (Wetter et al., 2012). Phase 7, which involved the final questionnaire, was invariably the concluding phase.

Phase 1: Welcoming and General Introduction. The moderator initiated the session by welcoming participants and verifying their access to the Decision Support System (DSS) and accompanying materials. A pre-recorded video was presented, offering a comprehensive overview of the DSS's features and fundamental operations.

Phase 2 (Training): Training Session and Usability Evaluation. Participants engaged in a series of tasks aimed at acquainting them with the DSS and allowing for system customization based on individual preferences. This phase incorporated interactive tutorials built into the DSS and practice exercises to facilitate navigation and feature utilization within the DSS. This phase was additionally designed to allow participants the opportunity to tailor the DSS settings according to their preferences. Upon completion, participants were required to fill out a questionnaire assessing usability and user experience.

Phase 3 (QC,BB): Manual Arousal Scoring and Re-Assessment with Black Box AI Assistance. Participants conducted manual arousal scoring on a new patient record, establishing a baseline for subsequent comparison. They then re-evaluated the same record using the DSS with black box AI assistance, which provided event suggestions without supplementary explanations. This setup enabled participants to adjust their manual scoring based on the AI-generated suggestions.

Phase 4 (QC,WB): Manual Arousal Scoring and Re-Assessment with Transparent AI Assistance. Participants performed manual arousal scoring on a new patient record, creating a baseline for later comparison. Subsequently, they re-assessed the same record using the DSS with transparent AI assistance. The system offered explanations and visualizations alongside its predictions, including confidence scores, decision thresholds, and both local and global feature importance for each proposed arousal event onset (refer to Section 3.4). This allowed participants to adjust their manual scoring based on the AI-generated suggestions.

Phase 5 (Start,BB): Arousal Scoring with Black Box AI Assistance. Participants conducted the arousal scoring task on a new patient record using the DSS with black box AI assistance, which provided event suggestions devoid of additional explanations.

Phase 6 (Start,WB): Arousal Scoring with Transparent AI Assistance. Participants utilized the DSS to perform arousal scoring with transparent AI assistance, which offered explanations and visualizations alongside its predictions. The system offered explanations and visualizations alongside its predictions, including confidence scores, decision thresholds, and both local and global feature importance for each proposed arousal event onset (refer to Section 3.4).

Phase 7: Final Questionnaire. In the concluding phase, participants completed a final questionnaire to evaluate their overall experience with the DSS, the perceived utility of the system, and their experiences in working with the AI-assistance. They were also encouraged to provide feedback and suggestions for potential improvements.

3.6.4 Provided Resources

Participants were provided with the following resources:

Written Instructions Comprehensive written instructions for the user study.

Moderator A moderator was present throughout the study to provide guidance.

Introductory Video An introductory video that showcased the primary features of the Decision Support System (DSS) and its AI assistance capabilities.

Interactive Tour An interactive tour integrated into the DSS.

Questionnaires A series of questionnaires was administered to gather participant feedback.

3.7 Evaluation Approach and Rationale

The evaluation approach is illustrated in Figure 6.

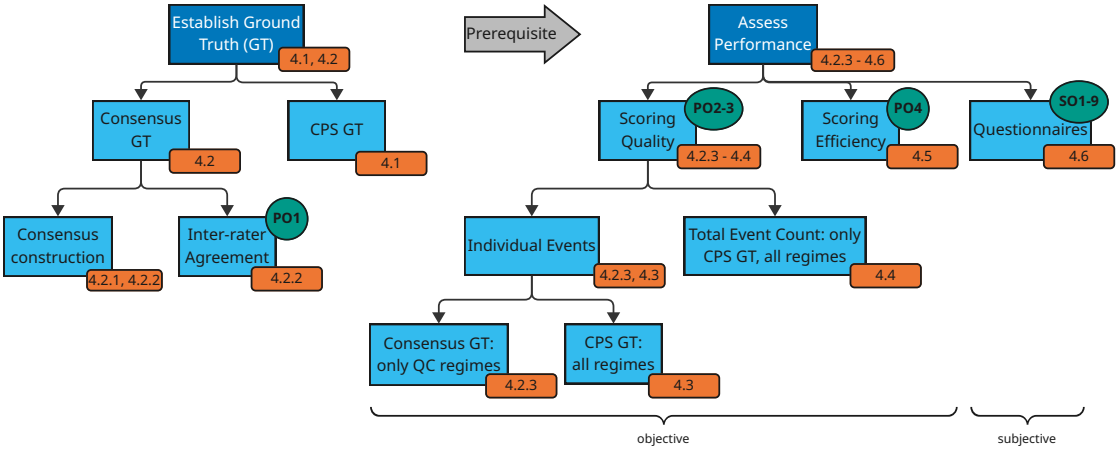


Figure 6: **Illustration of the Evaluation Process.** Sections displaying the results of each step are marked in orange, while the objectives are marked in teal. Abbreviations used: *CPS* denotes the Comprehensive Polysomnography dataset (Kraft et al., 2024) (see Section 3.5), *QC* stands for Quality Control, *PO* indicates Primary Objective, and *SO* means Secondary Objective (see Section 3.6.1).

Establishing an adequate ground-truth (GT) is the initial task for our evaluation which is of critical importance (Mohseni et al., 2021). We will utilize two variants: the *consensus GT* constructed from unaided scoring results of the study participants and the *CPS GT* from the CPS dataset. During this process, we also examine the inter-rater agreement among study participants (*PO1*), which is instrumental in constructing the consensus GT.

Having a GT specifically for human solo performance is essential. Merely comparing different explanation techniques is insufficient, as it does not provide insights into the utility of explanations versus no explanations (Rong et al., 2022).

To address the primary objectives *PO2-PO3*, we conduct objective performance evaluations. These evaluations are performed at two levels: per-event and total event count.

The *QC* scoring phases serve a dual purpose. Firstly, they allow us to evaluate the timing of AI assistance (*PO3*) by comparing it with the *Start* scoring phases. Secondly, they enable us to use the manual scoring baseline to construct the consensus GT. This construction helps us assess the unaided inter-rater agreement (*PO1*) and compare the performance of human-AI collaboration with human solo performance (*PO2*). Consequently, the *QC* phases can only be evaluated using the consensus GT (refer to Figure 6).

The secondary objectives *SO1-SO9* (see Section 3.6.1) are assessed through participants' subjective experiences. These are gathered using questionnaires that include both quantitative responses, such as Likert scale ratings, and qualitative responses in the form of open-ended questions. Thematic analysis will be applied to open-ended responses to extract deeper insights and identify areas for improvement.

By integrating objective performance metrics with subjective questionnaire responses, such as trust or perceived utility, we aim to provide a comprehensive evaluation of each objective and the overarching research questions. This discussion is carried out in Section 5.

3.8 Data Recording and Privacy

Participants are assured of anonymity throughout the study. In addition to collecting scoring data and questionnaire responses, the study involves screen and audio recordings, to which participants have consented. However, to maintain participant anonymity, these recordings will not be disclosed.

4 Results

This section presents the findings of the user study, addressing both the primary objectives (*PO1–PO4*) and secondary objectives (*SO1–SO9*) as defined in Section 3.6.1. General considerations about the evaluation approach are discussed in Section 3.7 and the structure is visualized in Figure 6.

We begin by outlining the process of ground truth selection in Section 4.1, where both the consensus ground truth and the CPS ground truth are introduced. Section 4.2 begins by analyzing the inter-rater agreement among study participants (*PO1*), which is fundamental to the development of the consensus ground truth. Building on this, the section further investigates how human solo annotators, the AI system, and human-AI teams perform when evaluated against the consensus ground truth. In contrast, Section 4.3 extends this performance comparison to the CPS ground truth, enabling a thorough assessment of each approach relative to both reference standards (*PO2*) and it further investigates how different timing and transparency modes affect the alignment of human-AI teams (*PO3*). In contrast to the preceding event-level analyses, Section 4.4 extends the evaluation of *PO2* and *PO3* on the CPS ground truth by shifting focus to aggregate arousal counts, providing a clinically oriented perspective on how human, AI, and human-AI teams compare in terms of total event detection across different collaboration regimes. An analysis of time efficiency (*PO4*) follows in Section 4.5. Finally, Section 4.6 presents the questionnaire results, evaluating the usability and user experience of the web application, providing additional insights into *PO2* and *PO3*, and addressing the secondary objectives (*SO1–SO9*) related to AI assistance. Furthermore, a comparison of feature attributions between DeepLIFT and GradientSHAP is presented in Appendix B.

4.1 Establishing the Ground Truth and Research Goals

A central challenge in evaluating performance for the arousal annotation task is the establishment of an appropriate ground truth. In this study, we employ two distinct reference standards: (1) the *CPS ground truth*, which comprises annotations from multiple human experts, with each patient recording scored by a single expert as part of routine clinical practice in the Comprehensive Polysomnography (CPS) dataset; and (2) a *consensus ground truth*, derived from manual, unaided scoring results of user study participants through clustering and probabilistic weighting.

The CPS ground truth reflects the cumulative expertise of about five medical professionals who contributed to the dataset over time during routine clinical practice. While each patient record is annotated by a single expert, the dataset as a whole incorporates the practices of several individuals. This reference standard was also used to train the AI model, potentially predisposing the model to replicate its inherent patterns.

In contrast, the consensus ground truth is designed to mitigate individual biases by synthesizing the judgments of multiple annotators into a unified standard. Specifically, it is constructed by combining the annotations for the two subjects initially scored manually without AI assistance (subjects S2 and S4 in Table 1; see also Section 3.6.3). This process includes annotations from all study participants as well as the CPS ground truth annotations. The resulting consensus, formed via clustering and expectation-maximization weighting (cf. Section 4.2), provides a more balanced and arguably less biased reference than the CPS ground truth.

Accordingly, we assess the performance of humans, AI, and human-AI teams from two perspectives. First, we compare results against the consensus ground truth, which is less likely to favor any

single approach and is thus more suitable for claims of objective correctness. Second, we evaluate performance relative to the CPS ground truth, which represents a distinct, albeit potentially more biased, standard. This dual approach enables us to examine whether human-AI teams tend to align more closely with a given standard (the CPS ground truth) compared to unaided human annotators, and how this alignment differs when evaluated against a neutral consensus reference. For the purposes of statistical inference, we designate the CPS ground truth as the definitive reference for addressing the primary objectives (*PO2* and *PO3*). In contrast, we employ the consensus ground truth to conduct sensitivity analyses, as detailed in Section 2.5.

4.2 Performance Under the Consensus Ground Truth

4.2.1 Clustering Manual Human Annotations

When employing the consensus as the evaluation standard, our objective is to assess the extent to which human solo annotators, the AI system, and human-AI teams align with a collectively derived reference for correct annotations. This approach is grounded in the premise that, in cases where true correctness is ambiguous, a consensus among multiple human experts can serve as a balanced ground truth.

Given the imperfect alignment among human annotations, we first cluster these annotations to delineate regions of agreement and disagreement. Recognizing that different clustering algorithms may yield varying results, we selected two representative and conceptually distinct methods for comparison: DBSCAN, a density-based clustering algorithm, and agglomerative clustering, a hierarchical approach (cf. Section 2.3.1).

For both clustering methods, we set the minimum cluster size to $m^{\text{Cluster}} = 2$, requiring that at least two annotators mark a temporal region for it to be considered a cluster. This conservative threshold is justified by the subsequent probabilistic weighting step, which enables us to filter out unlikely clusters of size two or greater.

Our DBSCAN-based clustering approach is described in Section 2.3.1. The key parameters to determine are the epsilon value ε and the minimum cluster size m^{Cluster}.

To select ε, we apply the Kneedle algorithm to the 1-distance graph for multiple sensitivity values S (see Section 2.3.1). This multi-parameter strategy produces a range of candidate ε values, facilitating manual selection informed by domain expertise and visual inspection of clustering outcomes. The results of this analysis are presented in Figure 7.

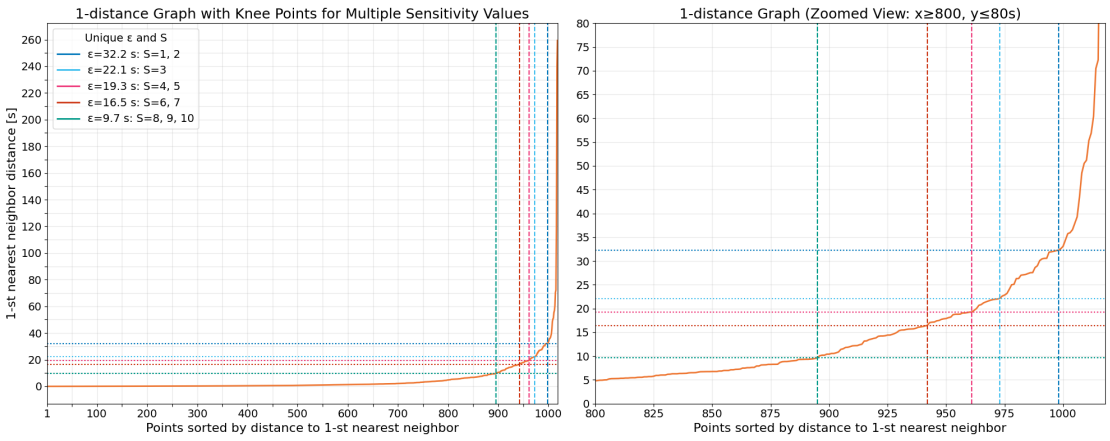


Figure 7: **1-Distance Graph with Knee Points for Multiple Sensitivity Values.** The left graph displays the 1-distance values sorted by distance to the 1st nearest neighbor, with knee points identified for sensitivity values $S \in \{1, \ldots, 10\}$. The right graph provides a zoomed view of the relevant range. These knee points inform the selection of the optimal epsilon parameter ε for DBSCAN clustering, offering a spectrum of candidate values for robust parameterization.

Because the ε parameter in DBSCAN does not impose a strict upper bound on cluster size, allowing clusters to span multiple ε distances, we select a relatively low value of $\varepsilon = 9664$ ms, as indicated by sensitivities $S \in \{8, 9, 10\}$. Here, ε is best interpreted as the minimum gap between two clusters. Notably, this value aligns well with the American Academy of Sleep Medicine guidelines (Berry et al., 2012), which stipulate a minimum of 10 seconds of sleep between arousals.

For agglomerative clustering, we utilize the Scikit-learn implementation and specify a distance threshold ε above which clusters are no longer merged, using the Ward linkage criterion. As this threshold is absolute, it is logical to base it on the typical duration of an arousal event, estimated by Boselli et al. (1998) at 14.6 ± 2.5 s, irrespective of age. This is consistent with the 15-second temporal tolerance buffer applied in our previous work (Kraft et al., 2025) for matching predicted and ground-truth events. Given that this is a statistical average, we set the threshold slightly higher at $\varepsilon = 19251$ ms, which remains within the 95% confidence interval of the arousal duration estimate.

In accordance with the minimum cluster size criterion ($m^{\text{Cluster}} = 2$), all annotations from a single annotator are assigned to a noise cluster (see Section 2.3.1).

4.2.2 Deriving the Consensus Clusters

To identify consensus clusters, i.e. those most likely to represent true events, it is instructive to first examine the degree of inter-rater agreement for the different clustering approaches.

Inter-Rater Agreement. Table 2 reports Krippendorff's Alpha, Fleiss' Kappa, and the mean pairwise Cohen's Kappa for both clustering methods.

Table 2: Inter-rater Agreement for Different Clustering Methods

Clustering Method	Krippendorff's Alpha	Fleiss' Kappa	Mean Pairwise Cohen's Kappa
DBSCAN	0.11	0.11	0.12
Agglomerative Clustering	0.09	0.09	0.10

All three metrics are marginally higher for DBSCAN than for agglomerative clustering, suggesting slightly greater annotator agreement within DBSCAN-derived clusters. Nevertheless, the absolute values remain low for both methods, indicating only minimal agreement above chance among annotators. To further investigate the inter-rater agreement, we examine pairwise agreement, as visualized in Figure 8.

The heatmaps in Figure 8 reveal two key findings. First, the overall agreement patterns are similar across both clustering methods. Second, despite generally low agreement, certain annotator pairs demonstrate fair (κ between 0.21 and 0.40) or even moderate (κ between 0.41 and 0.60) agreement. Concretely, in the DBSCAN results, fair agreement is observed for six annotator pairs (1&3, 1&7, 2&4, 4&7, 7&8, 7&9), and moderate agreement for one pair (3&7). Conversely, several annotator pairs exhibit very low, or even below-chance, agreement, most notably involving annotator 5. These findings indicate that, despite low overall agreement, there remains potential to derive a meaningful consensus from the human annotations.

To more rigorously assess sensitivity of the inter-rater agreement statistics, we examine how they vary as subsets of annotators are systematically excluded from the analysis. Let $n = 9$ be the total number of annotators and let each unit correspond to a clustered event. For each removal size $k \in \{1, \ldots, 7\}$ we enumerate all $\binom{n}{k}$ combinations of removed annotators. For a given combination, we recalculate Krippendorff's Alpha, Fleiss' Kappa, and mean pairwise Cohen's Kappa between the remaining $m = n - k$ annotators. The same three statistics computed with all n annotators serve as the baseline. The results of this analysis are presented in Figure 9.

Across removal sizes the medians remain close to the baseline, while the spread widens markedly for larger k. This pattern is expected: as the number of remaining raters $m = n - k$ decreases, inter-rater agreement estimates become more variable across subsets. Importantly, absolute levels across different k are not strictly comparable as each k corresponds to a different m. Since Alpha- and Kappa-based statistics are chance-corrected proportions, their magnitude depends not only on

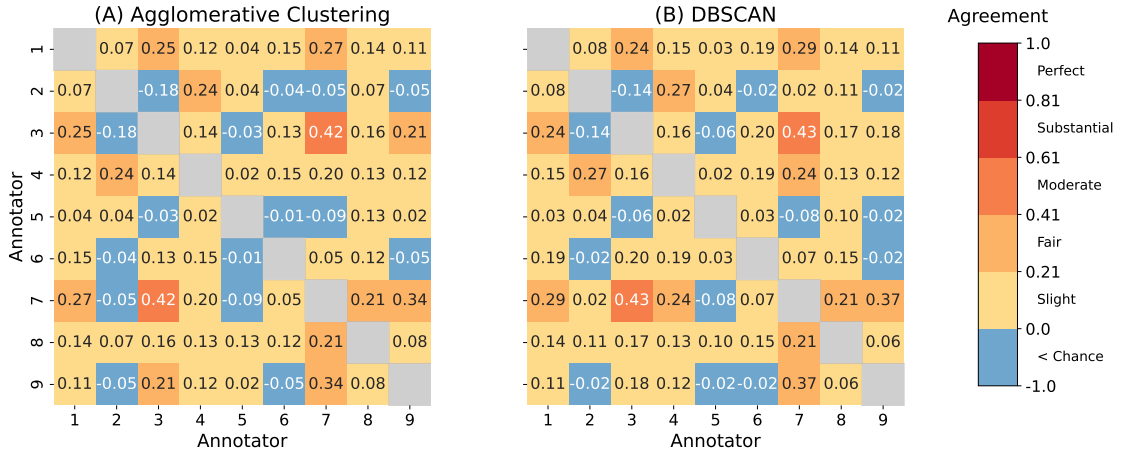


Figure 8: **Pairwise Cohen's Kappa Heatmaps** for the different clustering methods: (A) Agglomerative Clustering and (B) DBSCAN. The color scale indicates the level of agreement, ranging from perfect agreement (dark red) to less than chance agreement (light blue). Annotator 9 corresponds to the human annotations which served as the CPS ground truth for training the AI model.

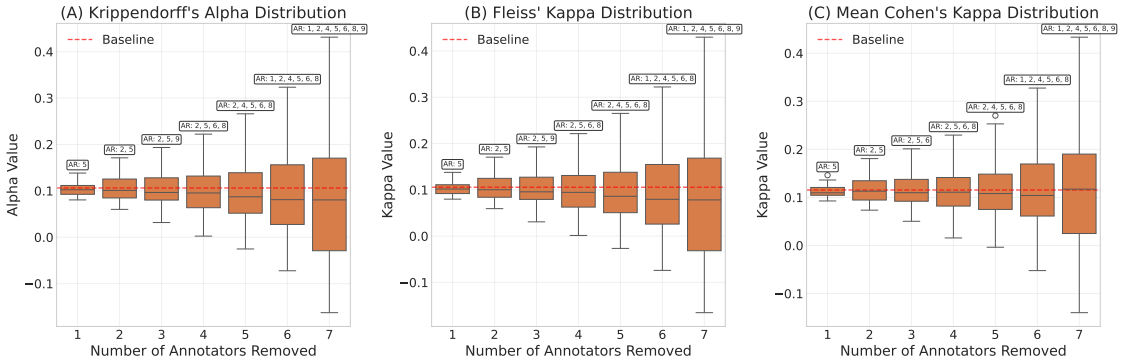


Figure 9: **Sensitivity of inter-rater agreement to systematic annotator removal.** For each $k \in \{1, \ldots, 7\}$, box plots show the distribution of Krippendorff's α (A), Fleiss' κ (B), and mean pairwise Cohen's κ (C), computed across all $\binom{9}{k}$ possible subsets formed by removing k annotators, with the unit (event) set fixed. The red dashed line marks the baseline value when all $n = 9$ annotators are included. The labels annotated above the boxplots indicate the particular set of annotators whose removal yielded the highest value for each agreement statistic ("AR" stands for "annotators removed"). Thus, while each box shows the spread of agreement for all combinations, only the set achieving the maximum is called out. As k increases, the variability in agreement widens, reflecting both reduced rater numbers and differences across subsets.

who remains but also on m and the label marginals. Within a fixed k, however, the interquartile range reveals meaningful heterogeneity and supports our former analysis: some subsets of annotators yield substantially higher agreement than others. Notably, removing annotators 5 and 2 leads to a marked increase in agreement across all three statistics, while the highest overall agreement is achieved between annotators 3 and 7, a pattern that aligns with the observations from Figure 8.

Nevertheless, to avoid introducing selection bias, we opted not to exclude any annotators from the consensus construction. Instead, we employ a probabilistic weighting scheme based on annotator agreement, utilizing an expectation-maximization algorithm (see Section 2.3.2).

This approach enables estimation of each annotator's sensitivity and specificity, providing a measure of trustworthiness that informs the consensus annotation process. The analysis is conducted separately for DBSCAN and agglomerative clustering, as illustrated in Figure 10.

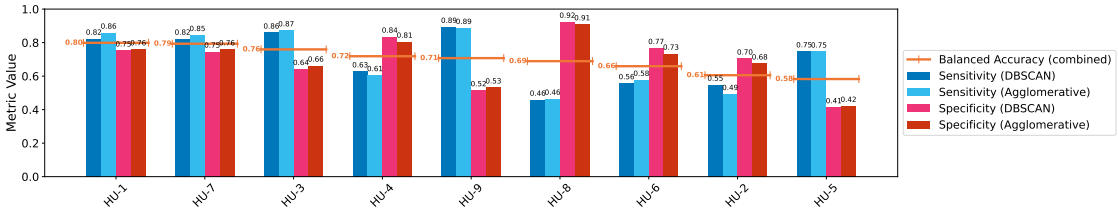


Figure 10: **Annotator Sensitivities, Specificities, and Balanced Accuracies for DBSCAN and Agglomerative Clustering.** For each annotator, sensitivities (blue bars) and specificities (red bars) are shown for both DBSCAN and agglomerative clustering. Annotators are sorted by their overall balanced accuracy, which is defined as the average of sensitivity and specificity for each clustering method, then averaged across both methods. The orange line indicates the balanced accuracy per annotator, highlighting variability in annotation performance. Values for both clustering approaches are closely aligned for most annotators, suggesting robustness of these performance metrics to the choice of clustering method.

Both clustering methods yield broadly similar outcomes, yet they also highlight pronounced differences in annotator reliability as reflected by inter-rater agreement. Notably, annotators 1, 7, and 3 achieve the highest balanced accuracy, underscoring their overall trustworthiness. In contrast, annotator 5 stands out for markedly low specificity and the lowest balanced accuracy, signaling a persistent pattern of unreliability relative to its peers. This low specificity reduces annotator 5's influence when voting against a cluster being an arousal event. However, due to high sensitivity, this annotator may still contribute meaningfully to identifying true arousal events. Notably, these results are in close agreement with the sensitivity analysis shown in Figure 9: retaining annotators 1, 3, and 7 produced the highest agreement when five annotators were excluded, while removing annotator 5 led to a notable improvement in agreement. This convergence reinforces the robustness of our findings across analytical approaches.

The sensitivities and specificities from Figure 10 are now incorporated as weights in the consensus voting scheme.

Figure 11 presents all human annotations, their DBSCAN-based clustering structure, the resulting consensus annotations, and the AI-generated annotations. A corresponding visualization for agglomerative clustering is provided in Appendix D.

Several observations merit emphasis. First, all annotators consistently agree on periods without annotations, which likely correspond to wakefulness, as the American Academy of Sleep Medicine guidelines (Berry et al., 2012) restrict arousal annotation to sleep periods. Second, distinct annotation *styles* are evident: some annotators (e.g., 2, 4, and 8) annotate sparsely, while others (e.g., 3, 7, and 9) annotate more densely. This heterogeneity may explain the presence of higher pairwise agreements that do not translate into high overall agreement. Finally, the AI annotations visually align closely with those of annotator 9, who provided the CPS ground truth for model training. Notably, the AI system appears to produce few false negatives but a relatively higher number of false positives compared to the consensus annotations, which is consistent with expectations, as it

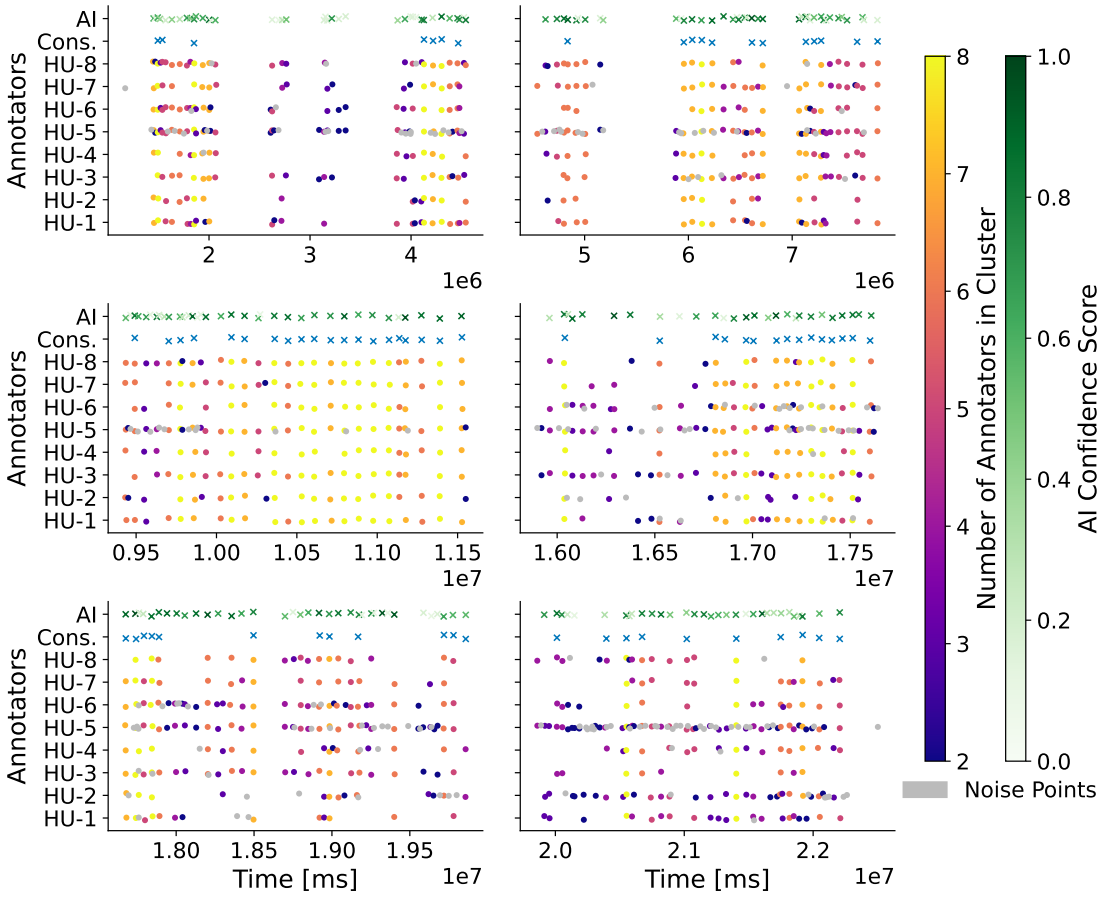


Figure 11: **Faceted Timeline Plot of Clustering Results using DBSCAN Clustering**. The plot displays human solo annotations across multiple panels, with data from patient samples *S2* and *S4* concatenated into approximately six hours of recording time. Annotations of the same color and vertical alignment belong to the same cluster. The horizontal displacement is for visual clarity only. Large intervals without annotations have been omitted. Consensus annotations (*Cons.*) are marked as blue crosses, while AI annotations are marked as green crosses, with the hue representing the AI model's confidence score. The color scale on the right indicates the number of annotators in each cluster. Annotator *HU-9* represents the human annotations used as the CPS ground truth for AI model training.

was optimized for the F2-score.

Comparison of Clustering Approaches. To assess the reliability and suitability of the consensus event clusters derived from different clustering algorithms, we compare the results of agglomerative clustering and DBSCAN using statistics and pairwise clustering similarity measures. The results are summarized below.

Figure 12 shows statistics for the clustering approaches.

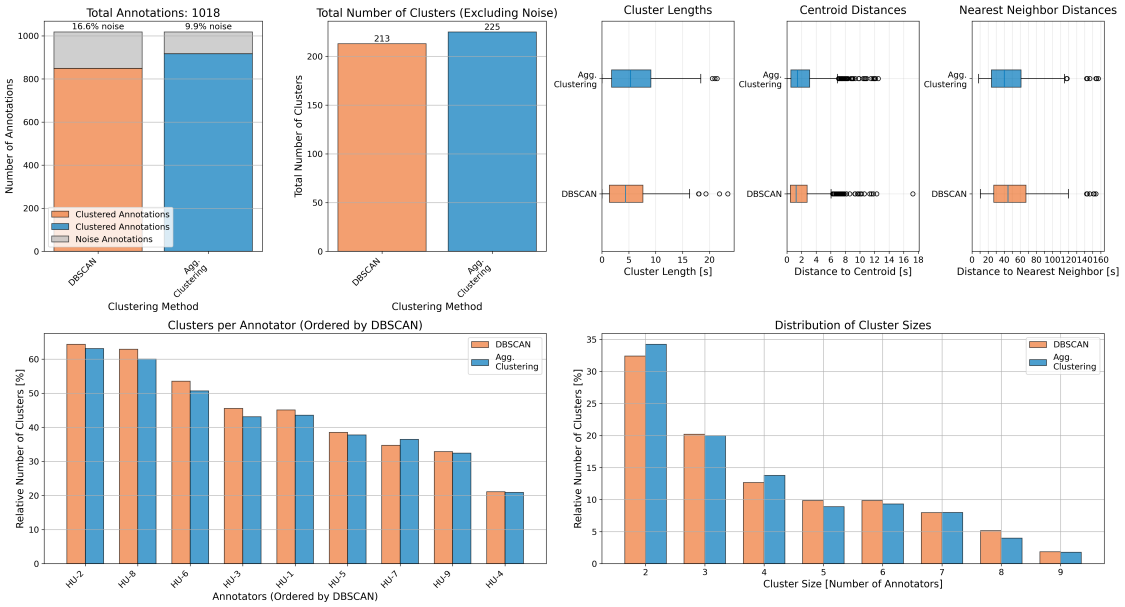


Figure 12: **Clustering Statistics** for DBSCAN and Agglomerative Clustering methods, illustrating various cluster characteristics. From top left to bottom right: Number of clustered versus noise annotations, total number of clusters, cluster length distributions, distances of annotations to their respective cluster centroids or the centroids of their nearest neighbor cluster, relative number of clusters each annotator is part of, and the distribution of cluster sizes. The statistics for both clustering methods are similar, suggesting that the choice of clustering method does not significantly impact the cluster structure.

The analysis highlights several distinctions between DBSCAN and agglomerative clustering. Notably, DBSCAN classifies a greater number of annotations as noise, thereby focusing attention on more meaningful clusters (top left). It also produces slightly fewer, yet more compact, clusters (top right), which may enhance the delineation of cluster boundaries. Furthermore, DBSCAN tends to incorporate a larger number of annotators per cluster (bottom left and bottom right), potentially reflecting a broader consensus among annotators. Despite these differences, both methods yield clusters with broadly similar characteristics, underscoring the robustness of the consensus clusters irrespective of the clustering algorithm employed.

This conclusion is further supported by pairwise similarity measures between the clusterings, as shown in Table 3.

The results indicate strong to very strong agreement between the clustering methods, as reflected by the Adjusted Rand Index (ARI) and Normalized Mutual Information (NMI) metrics. Specifically, the direct comparison between DBSCAN and agglomerative clustering yields an ARI of 0.57, signifying strong agreement, and an NMI of 0.95, indicating very strong agreement. This suggests that both algorithms identify similar cluster structures, with only minor differences in the assignment of individual data points.

When comparing the consensus clusters derived from each method, the agreement becomes even more pronounced, with an ARI of 0.90 and an NMI of 0.95, both indicative of very strong concor-

Table 3: **Pairwise Measures for Clustering Methods DBSCAN and Agglomerative Clustering**, including Adjusted Rand Index (ARI) and Normalized Mutual Information (NMI). The initial comparison is between the two clustering methods, the second comparison is between the consensus clusters derived from each method.

Comparison	ARI	NMI
Initial	0.57	0.95
Consensus	0.90	0.95

dance. This heightened consistency underscores the effectiveness of the consensus-building approach in capturing the underlying structure of the data. The consensus process not only enhances agreement between methods but also demonstrates that the resulting cluster structure is robust and largely independent of the specific clustering algorithm chosen.

Rationale for Selecting DBSCAN Consensus Clusters. Our analysis demonstrates that both DBSCAN and Agglomerative Clustering reliably identify well-defined, natural groupings within the data, as evidenced by consistently high values of Normalized Mutual Information (NMI) and Adjusted Rand Index (ARI). These metrics indicate that the choice of clustering algorithm does not substantially impact the resulting consensus events, and the consensus-building process further improves consistency between methods by reducing noise and enhancing cluster quality. The stability of the fundamental cluster structure, reflected in the high NMI values, provides confidence that the clustering results capture genuine patterns in the annotation data rather than artifacts of algorithm selection. Nevertheless, DBSCAN offers specific advantages that render it particularly well-suited to our application, motivating our choice to adopt DBSCAN consensus clusters for subsequent analyses.

Enhanced inter-rater agreement: DBSCAN yields higher inter-rater reliability metrics—including Krippendorff's Alpha, Fleiss' Kappa, and mean pairwise Cohen's Kappa—signifying improved clustering quality and greater annotator consensus.

Superior cluster properties: By discarding a greater proportion of annotations as noise, DBSCAN produces fewer and more compact clusters. This not only sharpens cluster boundaries but also increases the number of annotators represented within each cluster, reflecting stronger expert agreement.

Accordingly, we adopt the consensus clusters generated by the DBSCAN algorithm for all subsequent analyses.

4.2.3 Human and AI Solo Performance

Using the consensus ground truth, we evaluate the performance metrics of human solo annotators (HU), human-AI teams (HU&AI), and the AI model.

Distance Threshold Selection. As outlined in Section 2.4.1, it is necessary to define a distance threshold to determine whether an annotation should be considered part of a consensus cluster or classified as noise. Figure 12 (top right) presents the distribution of distances from human solo annotations to the centroids of their assigned consensus clusters, as well as to the centroids of their nearest neighboring clusters. The minimum distance to the nearest neighbor cluster (distinct from the assigned cluster) is approximately 10 seconds, which closely aligns with the chosen DBSCAN parameter of $\varepsilon = 9664$ ms. Conversely, only a small number of outlier annotations are located farther than ε from their own cluster centroid. Consequently, we adopt ε as the distance threshold for classifying annotations as either part of a consensus cluster or as noise. For human solo annotations, this approach results in only a few annotations being incorrectly labeled as noise, while effectively preventing misattribution of annotations to incorrect clusters.

Results. The corresponding results are summarized in Table 4.

Table 4: **Performance comparison with consensus ground truth**. Performance metrics for each human annotator (HU-n), human-AI team (HU-n&AI) which are aggregated across black box and white box AI assistance for quality control (i.e., phases 3 and 4 as described in Section 3.6.3), and the AI model. HU and HU&AI denote the averages across all human annotators and human-AI teams, respectively. HU-9 corresponds to the human annotations used as the CPS ground truth for AI model training.

Expert	**F1 Score**	**F2 Score**	**Precision**	**Recall**	**TP**	**FP**	**FN**
HU-1	0.70	0.76	0.61	0.82	62	40	14
HU-2	0.46	0.50	0.40	0.53	40	59	36
HU-3	0.65	0.76	0.52	0.86	65	59	11
HU-4	0.62	0.62	0.62	0.62	47	29	29
HU-5	0.39	0.54	0.26	0.74	56	156	20
HU-6	0.48	0.51	0.43	0.54	41	54	35
HU-7	0.69	0.75	0.60	0.80	61	41	15
HU-8	0.56	0.50	0.73	0.46	35	13	41
HU-9 (CPS GT)	0.56	0.71	0.41	0.87	66	94	10
HU Avg.	0.57	0.63	0.51	0.69	53	61	23
HU-1&AI	0.61	0.73	0.48	0.84	64	69	12
HU-2&AI	0.45	0.53	0.37	0.59	45	78	31
HU-3&AI	0.62	0.75	0.48	0.88	67	73	9
HU-4&AI	0.63	0.71	0.53	0.78	59	53	17
HU-5&AI	0.39	0.61	0.24	0.99	75	236	1
HU-6&AI	0.56	0.72	0.41	0.88	67	95	9
HU-7&AI	0.68	0.78	0.55	0.87	66	53	10
HU-8&AI	0.62	0.75	0.49	0.86	65	67	11
HU&AI Avg.	0.57	0.70	0.44	0.84	64	90	12
AI	0.51	0.71	0.35	0.96	73	137	3

On average, human solo annotators (HU) achieved an F1 score of approximately 0.57, while the AI model alone attained an F1 score of 0.51.

A one-sample t-test (see Table 5) comparing the AI's F1 score to the distribution of human F1 scores did not reject the null hypothesis of equal performance ($p = 0.15$), indicating no statistically significant difference between the AI model and human solo annotators.

Table 5: **Event-level performance (F1 score) evaluated against the consensus ground truth.** Statistical methods are described in Section 2.5. Here, n denotes the number of participants. p-values are calculated using exact sign-flip tests, and 95% confidence intervals (CIs) are obtained via percentile bootstrap. Across all comparisons, the null hypothesis of equal performance cannot be rejected.

Comparison (A vs B)	n	A mean	B mean	Diff (A−B) [95% CI]	t(df)	p_{perm}
HU+AI vs HU	8	0.57	0.57	0.0031 [−0.030, 0.037]	0.17 (7)	0.92
HU+AI vs AI	8	0.57	0.51	0.06 [−0.0079, 0.12]	1.7 (7)	0.16
HU vs AI	9	0.57	0.51	0.06 [−0.019, 0.12]	1.6 (8)	0.15

Several factors should be considered when interpreting these findings. The AI model was optimized for the F2 score, as discussed in Kraft et al. (2025), under the assumption that this would be advantageous for human-AI collaboration compared to optimizing for the F1 score. As a result, the AI's solo performance is somewhat disadvantaged when evaluated using the F1 metric. The potential for a higher F1 score is evident in Figure 13. The left panel displays F1 and F2 scores across a range of decision thresholds, including values higher than the threshold used in the study (0.11), which was selected to maximize the F2 score on the training data. The threshold that maximizes the F1 score is 0.29. Applying this threshold to the AI model on the consensus ground truth (right panel) increases the F1 score from 0.51 to 0.59, slightly surpassing the average human solo performance and further supporting the conclusion of no significant difference between AI and human annotators.

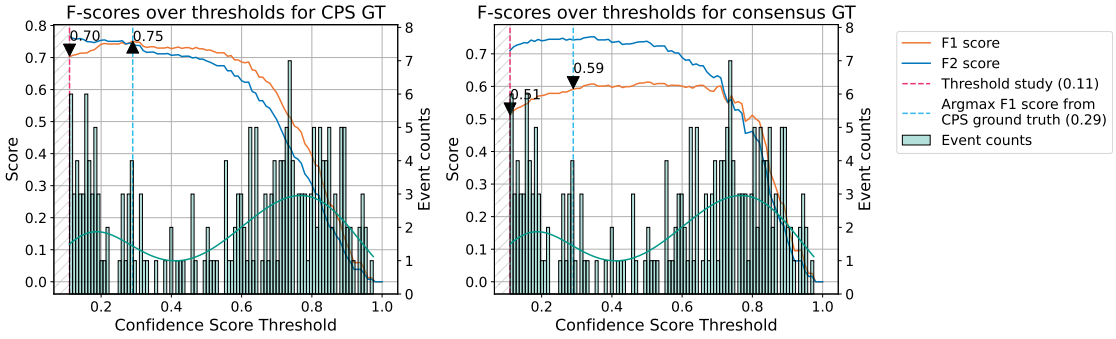


Figure 13: **F1 and F2 Scores over Confidence Score Thresholds** for both CPS ground truth (left) and consensus ground truth (right). The vertical lines indicate the threshold used in the study (0.11) and the threshold corresponding to the maximum F1 score on the CPS ground truth, which is also applied to the consensus ground truth. Maximum F1 score values at these thresholds are annotated. The distribution of event counts is also shown. The AI model demonstrates superior calibration on the CPS ground truth compared to the consensus ground truth.

Although, in practice, such a threshold would not be selected based on test data as in this illustrative example, these results demonstrate that optimizing the threshold for the F1 score on training data typically leads to improved F1 performance on the consensus ground truth as well. This observation underscores the argument that the AI model's performance can be further enhanced through appropriate calibration.

It is also important to note that the samples analyzed here (S2 and S4 in Table 1) are among those where the AI performed best in the test set. While human performance may also vary across

samples, the AI model may not generalize as robustly as human annotators to other, more challenging cases.

In summary, these findings indicate that the DeepSleep AI model, trained on annotations from multiple human scorers (one per sample), is capable of achieving human-level performance when evaluated against the consensus of expert annotators.

4.2.4 Human-AI Team Performance

We next examine the performance of human-AI teams relative to human solo performance, as assessed against the consensus ground truth annotations. In this setting, human scorers first completed manual scoring and were subsequently permitted to revise their annotations based on AI-generated suggestions for quality control (see phases 3 and 4 in Section 3.6.3).

As shown in Table 4, the incorporation of AI assistance does not, on average, enhance collaborative performance compared to human solo performance when measured by the F1 score. This finding is further supported by the paired t-test results in Table 5. Although recall improves with AI support, this benefit is offset by a reduction in precision. To better understand this outcome, we refer to Figure 14, which depicts the distributions of added and deleted events for human-AI teams compared to human solo performance across all annotators.

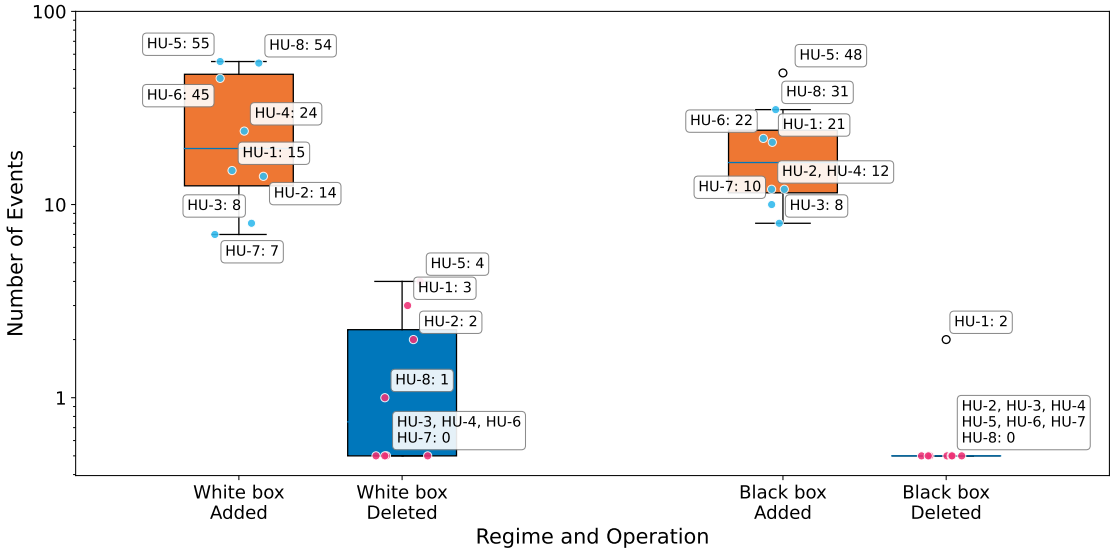


Figure 14: **Distribution of Added and Deleted Events** after humans went through an AI assisted quality control phase. This plot illustrates the number of events added and removed across different regimes (White box and Black box) for each human annotator (HU-n). The distributions are depicted on a logarithmic scale to enhance readability, given the substantial disparity between the number of added and deleted events.

The figure reveals a pronounced imbalance between the number of added and deleted annotations across regimes. In the white box regime, annotators added a substantial number of events, with HU-5 and HU-8 contributing most prominently. Conversely, annotation deletions were infrequent, with half of the annotators not deleting any events. In the black box regime, the tendency to add annotations persisted, though generally at a lower rate than in the white box regime, and only one annotator deleted any events. These patterns suggest that white box AI suggestions exert a stronger influence on annotators than black box suggestions. Moreover, black box AI support appears to further diminish the already limited inclination of scorers to critically reassess their own annotations compared to white box support.

Given that Table 4 demonstrates the AI model introduces a substantial number of false positives, we conclude that, when evaluated against the consensus ground truth, human annotators are gen-

erally unable to effectively leverage AI suggestions for quality control. On average, they are unable to reliably distinguish between correct and incorrect AI-generated annotations.

Finally, we highlight a notable outlier among the human annotators: HU-5. This annotator produced substantially more annotations than the other experts (see Table 4), leading to low pairwise Cohen's Kappa values in Figure 8 and the lowest specificity for consensus construction (see Figure 10). This in turn resulted in the lowest F1 score and precision among all experts (see Table 4). Interestingly, while AI assistance for quality control reduced HU-5's number of false negatives to the lowest among all experts, it also led to a substantial increase in false positives, since HU-5 predominantly added events suggested by the AI in the quality control phase (see Figure 14). Observationally, HU-5 was also the only expert who, during phase 5 of the study, chose to validate only the AI suggestions when annotating with black box AI assistance from the outset, effectively skipping all time steps not highlighted by the AI. This behavior indicates a particularly high level of trust and reliance on the AI model. We will further investigate this annotator in subsequent sections.

4.3 Performance Under the CPS Ground Truth

Having established that AI assistance does not enhance human scorer performance when evaluated against the consensus ground truth, we now turn to a related question: To what extent can human-AI teams align with the specific standard on which the AI model was trained, i.e. the CPS ground truth, compared to unaided human annotators?

The CPS ground truth serves as the reference standard for the CPS dataset, derived from routine clinical practice. For each subject, annotations were provided by a single medical expert, with the specific expert varying across subjects.

Although this ground truth is not entirely neutral, as it is reflecting the individual styles of the contributing scorers, it provides a concrete standard for assessing alignment. In the previous section, the CPS ground truth was represented by one human expert (*HU-9*). This perspective shifts the research question: Do human-AI teams conform more closely to a given standard than human annotators working independently?

4.3.1 Comparison of Human-AI Team, Human Solo, and AI Solo Performance

Table 6 presents the comparative results.

Both the AI model and the average human-AI team outperform the average human solo annotator, primarily due to higher recall, with a modest improvement in precision as well. The substantial performance gap between the AI model and the human solo average is statistically significant (see Table 7), which is expected given that the AI was trained on this distribution.

Table 2 further reveals only fair agreement between one participant (annotator 7) and the CPS ground truth (annotator 9), and slight agreement for four other participants (annotators 1, 3, 4, and 8).

A key question is whether human-AI team performance differs significantly from human solo performance. To address this, we conducted a paired t-test, with results shown in Table 7.

The p-value of 0.012 is below the significance threshold, providing evidence that human-AI team performance is significantly different from human solo performance when evaluated against the CPS ground truth. This finding stands in stark contrast to the results based on the consensus ground truth (Section 4.2), where no such improvement was observed. Here, AI-generated suggestions appear to guide human annotators toward the style or criteria embodied in the CPS ground truth.

These results indicate that, when the objective is to align with a specific established standard (such as the CPS ground truth used for AI training), AI assistance can indeed help human annotators move closer to that benchmark.

Detailed Analysis of Human-AI Team Performance The boxplot in Figure 15 offers further insight into the performance distributions of human solo annotators, human-AI teams, and the benefit ratio, which quantifies the effectiveness of AI integration (see Section 2.4.1).

Table 6: **Performance Comparison with CPS Ground Truth** used for AI Model Training. Performance metrics are reported for each human annotator (HU-n), human-AI team (HU-n&AI), and the AI model. HU and HU&AI denote the averages across all humans and all human-AI teams, respectively.

Expert	**F1 Score**	**F2 Score**	**Precision**	**Recall**	**TP**	**FP**	**FN**
HU-1	0.49	0.43	0.63	0.40	64	38	96
HU-2	0.36	0.32	0.47	0.29	47	52	113
HU-3	0.55	0.51	0.63	0.49	78	46	82
HU-4	0.39	0.32	0.61	0.29	46	30	114
HU-5	0.44	0.48	0.39	0.51	82	130	78
HU-6	0.32	0.28	0.43	0.26	41	54	119
HU-7	0.56	0.49	0.72	0.46	73	29	87
HU-8	0.29	0.22	0.62	0.19	30	18	130
HU Avg.	0.42	0.38	0.56	0.36	58	50	102
HU-1&AI	0.59	0.56	0.65	0.54	86	47	74
HU-2&AI	0.44	0.41	0.50	0.39	62	61	98
HU-3&AI	0.57	0.55	0.61	0.54	86	54	74
HU-4&AI	0.54	0.49	0.65	0.46	73	39	87
HU-5&AI	0.56	0.70	0.43	0.83	133	178	27
HU-6&AI	0.53	0.54	0.53	0.54	86	76	74
HU-7&AI	0.61	0.56	0.71	0.53	85	34	75
HU-8&AI	0.61	0.58	0.67	0.56	89	43	71
HU&AI Avg.	0.56	0.55	0.60	0.55	88	66	72
AI	0.70	0.76	0.62	0.81	130	80	30

Table 7: **Event-level performance (F1-score) evaluated against the CPS ground truth.** Pairwise comparisons follow the methodology outlined in Section 2.5, with each comparison based on $n = 8$ participants. Permutation p-values are computed using exact two-sided sign-flip tests, and 95% confidence intervals (CIs) are estimated via the percentile bootstrap method. No adjustment for multiple comparisons is necessary, as each primary question constitutes a single statistical family. As all human-AI teams outperform their respective human solo counterparts and the AI solo outperforms all others (see Table 6), the sign of all differences is the same, resulting in identical permutation p-values across comparisons. In all cases, the null hypothesis is rejected at the 5% significance level.

Comparison (A vs B)	A mean	B mean	Diff (A−B) [95% CI]	t(df)	p_{perm}
HU+AI vs HU	0.56	0.42	0.13 [0.075, 0.20]	3.9 (7)	0.012
HU+AI vs AI	0.56	0.70	−0.15 [−0.19, −0.11]	−7.4 (7)	0.012
HU vs AI	0.42	0.70	−0.28 [−0.34, −0.21]	−7.8 (7)	0.012

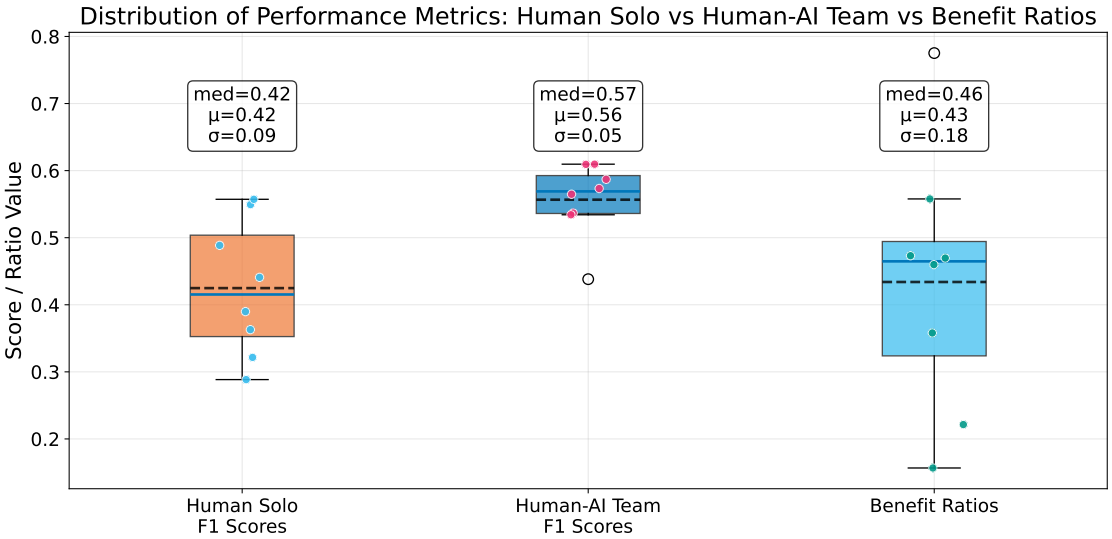


Figure 15: **Distribution of Performance Metrics:** F1 scores for human solo annotation, human-AI team collaboration, and the corresponding benefit ratios. Boxplots display the mean (black dashed line), interquartile range (box), whiskers, individual data points (colored circles), and outliers (empty circles). The data reveal that AI assistance substantially improves human performance and consistency, though benefit ratios vary considerably across teams.

The distribution of human solo F1 scores has a mean ($\mu = 0.42$) and standard deviation ($\sigma = 0.09$), reflecting notable variability in individual performance. The median is approximately 0.41, with the interquartile range (IQR) spanning roughly 0.36 to 0.50.

In contrast, human-AI team F1 scores exhibit a higher mean ($\mu = 0.56$) and a lower standard deviation ($\sigma = 0.05$), indicating that AI collaboration not only enhances performance but also yields more consistent results across teams. The median team F1 score is around 0.57, with a narrower IQR from approximately 0.53 to 0.60. This upward shift in both mean and consistency underscores the positive impact of AI assistance. An outlier near 0.43 suggests that, in at least one instance, a human-AI team performed closer to the human solo baseline.

The benefit ratio distribution has a mean ($\mu = 0.43$) numerically close to the human solo F1 mean, but its interpretation is distinct. Notably, the benefit ratio exhibits a much larger standard deviation ($\sigma = 0.18$) than either F1 distribution, indicating substantial variability in how effectively different teams leveraged AI support. The median benefit ratio is approximately 0.46, suggesting that, on average, teams reached about 46% of the potential alignment improvement offered by the AI. The presence of an outlier at approximately 0.78 indicates that at least one team was highly effective in adopting the AI's alignment, while the lower end of the distribution reflects cases with minimal benefit.

In summary, these findings demonstrate that human-AI collaboration generally results in higher and more consistent F1 scores compared to human solo annotation. However, the broad distribution of benefit ratios highlights considerable variability in the extent to which teams capitalize on AI assistance. This suggests that factors beyond the AI's intrinsic capabilities, such as the design of human-AI interaction, user training, or task complexity, may substantially influence the realized benefits of AI support.

4.3.2 Analysis of AI Assistance Regimes

As outlined in Section 3.6.3, we aggregated human-AI team performance across experimental phases into AI support regimes. Table 8 presents detailed outcomes for each regime, including both atomic and composite regimes (the latter formed by concatenating samples across phases which leads to

metric scores on the micro-level). The table reports the average human-AI team F1 score, the standalone AI F1 score, and the *relative F1 scores* $\mathcal{R} = \text{F1}_{\text{HU+AI Avg.}}/\text{F1}_{\text{AI}}$ (see also Section 2.4.1).

Table 8: **F1 score comparison across AI support regimes**, ordered by the relative F1 scores $\mathcal{R}$. Abbreviations: Start denotes AI support from the outset; QC indicates AI support used exclusively for quality control after initial manual scoring; BB and WB refer to black box and white box AI support, respectively. Regimes with a "+" are composite regimes, constructed by concatenating samples from the respective phases. All others are atomic regimes. $\text{F1}_{\text{HU+AI Avg.}}$ is the average human performance with AI support, F1_{AI} is the AI solo performance, and $\mathcal{R}$ are the relative F1 scores of $\text{F1}_{\text{HU+AI Avg.}}$ to F1_{AI}. Phases are detailed in Section 3.6.3. Values are means with 95% confidence intervals.

Regime	Phase	$\text{F1}_{\text{HU+AI Avg.}}$	F1_{AI}	$\mathcal{R}$
QC, WB	4	0.59 [0.53, 0.65]	0.66	0.90 [0.81, 0.98]
Start+QC, WB	4, 6	0.57 [0.53, 0.61]	0.65	0.87 [0.81, 0.94]
Start, WB	6	0.54 [0.50, 0.58]	0.64	0.84 [0.78, 0.89]
Start, BB+WB	5, 6	0.56 [0.52, 0.61]	0.70	0.80 [0.73, 0.86]
QC, BB+WB	3, 4	0.56 [0.51, 0.60]	0.70	0.79 [0.73, 0.86]
Start, BB	5	0.58 [0.51, 0.65]	0.74	0.78 [0.68, 0.88]
Start+QC, BB	3, 5	0.55 [0.49, 0.60]	0.75	0.73 [0.66, 0.80]
QC, BB	3	0.52 [0.47, 0.57]	0.75	0.69 [0.63, 0.75]

Analysis Overview We analyze the log ratio of relative F1 scores $\mathcal{R} = \text{F1}_{\text{HU+AI Avg.}}/\text{F1}_{\text{AI}}$ with a small offset ($\varepsilon = 10^{-8}$) to avoid zeros, following the statistical modeling approach described in Section 2.6.2.

Primary Analysis: RM-ANOVA with Planned Orthogonal Contrasts Following the methodology described in Section 2.6.3, we conduct RM-ANOVA and planned orthogonal contrasts. The results are summarized in Table 9.

The RM-ANOVA reveals a significant main effect of *AI* and a significant *AI*× *Timing* interaction, both associated with large effect sizes. In contrast, the main effect of *Timing* does not reach significance. The ANOVA employs parametric F-tests, while the planned contrasts use permutation tests with Holm correction, accounting for the different p-values between approaches.

Secondary Analysis: Planned-contrast t-tests We conduct planned-contrast tests for the main effects and their interaction, following the methodology described in Section 2.6.4. Given the significant interaction in the ANOVA, we examine *simple effects* (comparing WB vs. BB at both Start and QC, and Start vs. QC within both WB and BB conditions) using the approach outlined in Section 2.6.6. Results are presented in Tables 10 and 11.

Our findings indicate that white box (WB) assistance improves relative F1 scores compared to black box (BB) assistance by approximately **18%** on average (AI main effect), with a large within-subject effect size ($d_z = 1.38$), reflecting substantial enhancement in relative F1 performance. The interaction effect is also pronounced ($d_z = 1.15$ in absolute value), and the simple effects analysis reveals that the benefit of AI transparency is most pronounced during the QC phase. Specifically, at QC, WB outperforms BB by about **30%** on average, with a very large effect size ($d_z = 2.22$). In contrast, the difference between WB and BB at Start is small to moderate ($d_z = 0.45$) and not statistically significant. The timing difference within BB is large ($d_z = 1.06$), reaching significance prior to multiplicity correction, but not after familywise error control.

Table 9: **RM-ANOVA summary for main and interaction effects.** Each effect has $df_{\text{effect}} = 1$ and $df_{\text{error}} = 7$. Effect sizes are reported as partial eta-squared η_{p}^2. We find that the AI main effect and interaction effect are significant, while the timing main effect is not significant.

Effect	$F(1, 7)$	p	η_{p}^2
AI main (WB vs. BB)	**15.268**	**0.0058**	**0.686**
Timing main (Start vs. QC)	0.596	0.466	0.078
AI × Timing Interaction	**10.540**	**0.014**	**0.601**

Table 10: **Planned-contrast test results for main and interaction effects in the** 2×2 **repeated-measures design.** Each row reports the paired t-test statistic, permutation p-value with familywise error control via Holm correction (applied to the two main effects while the interaction uses raw permutation p), the estimated effect ratios (with bootstrap 95% confidence interval), and within-subject Cohen's d_z effect size (with bootstrap 95% confidence interval). These results align with the RM-ANOVA findings, demonstrating that white box (WB) AI assistance yields a statistically significant and substantial improvement in relative F1 performance compared to black box (BB) assistance (an average increase of approximately **18%**), and that the interaction effect is also statistically significant.

Contrast	$t(7)$	p_{perm}	Ratio [95% CI]	d_z [95% CI]
AI main	**3.91**	**0.039**	**1.18 [1.09, 1.27]**	**1.38 [0.59, 4.68]**
Timing main	0.77	0.46	1.03 [0.96, 1.08]	0.27 [–0.37, 2.47]
Interaction	**–3.25**	**0.027**	**0.83 [0.75, 0.92]**	**–1.15 [–2.51, –0.60]**

Table 11: **Simple-effect contrasts examining the significant interaction in the** 2×2 **repeated-measures design.** Each row reports a paired t-test for a specific simple effect and reports Holm-corrected permutation p-values. Effect sizes are given as within-subject Cohen's d_z with bootstrap-derived 95% confidence intervals. Effect ratios are accompanied by 95% bootstrap confidence intervals. We see that the benefit of white box (WB) AI assistance over black box (BB) is most pronounced during the quality control (QC) phase, with an average improvement of **30%**.

Simple effect	$t(7)$	$p_{\text{perm, Holm}}$	Ratio [95% CI]	d_z [95% CI]
WB vs. BB at Start	1.27	0.41	1.08 [0.96, 1.20]	0.45 [–0.23, 2.22]
WB vs. BB at QC	**6.27**	**0.047**	**1.30 [1.20, 1.40]**	**2.22 [1.57, 4.49]**
Start vs. QC with WB	–1.40	0.41	0.94 [0.86, 1.02]	–0.50 [–1.80, 0.19]
Start vs. QC with BB	2.99	0.11	1.13 [1.04, 1.20]	1.06 [0.31, 3.45]

4.4 Count-based evaluation of arousal scoring regimes

Rationale and link to objectives While event-level performance evaluation is valuable for understanding scorer behavior during the diagnostic arousal annotation task of polysomnographic patient data, clinical decision-making typically relies on more aggregated statistics, most notably, the arousal index, defined as the number of arousals per hour of sleep (Berry et al., 2012; Ehrlich et al., 2024). To address the primary objectives *PO2* and *PO3* (see Section 3.6.1) we therefore analyze absolute arousal counts against the CPS ground truth on three measures (see Section 2.4.2 for definitions): (i) *count accuracy* $A_{x,\mathrm{GT}}$, (ii) the *AI-Baseline Improvement Ratio* R_{GT} comparing the team to the AI (analyzed on $y_{\mathrm{RGT}} = \log R_{\mathrm{GT}}$), and (iii) the *percentage error* PE giving a measure of systematic count bias toward over- or under-counting. The results are visualized in Figure 16.

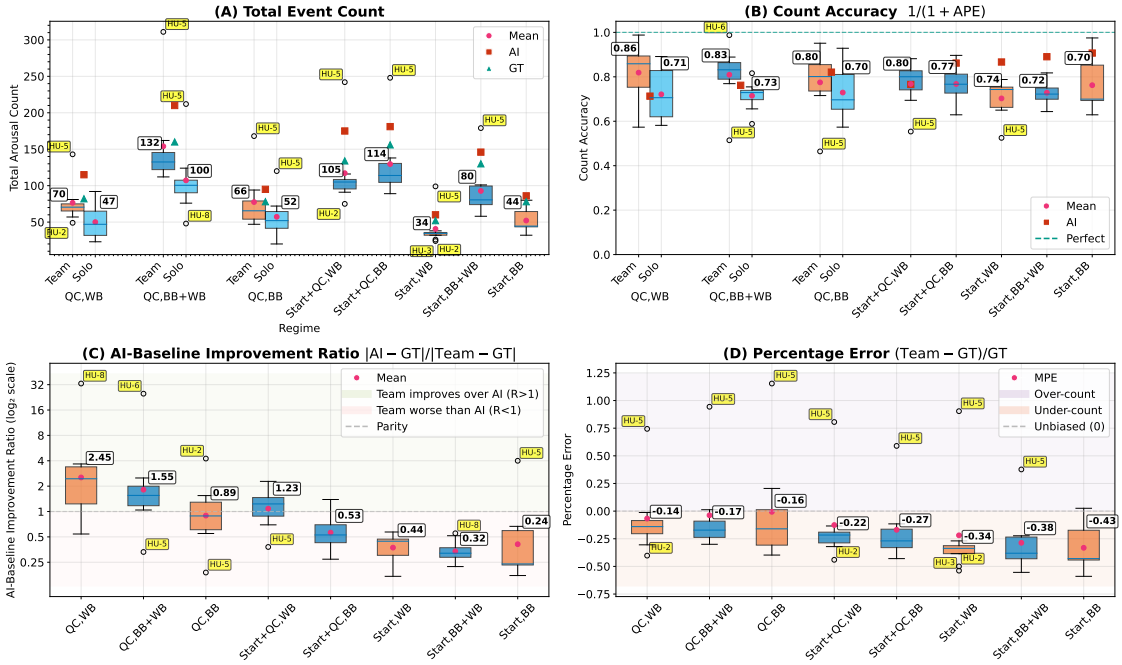


Figure 16: **Count-based evaluation across regimes.** Panel (**A**) displays the total arousal counts per recording for each scoring source: human solo scorers (C_{HU}, available only in the QC regime), human-AI teams ($C_{\mathrm{HU+AI}}$), AI-only outputs (C_{AI}, represented by red squares), and the ground truth (C_{GT}, represented by green triangles). Panel (**B**) presents the ground truth count accuracy $A_{x,\mathrm{GT}}$. Panel (**C**) illustrates the ground truth error reduction, visualized as the base-2 logarithm of the error reduction ratio, $\log_2 R_{\mathrm{GT}}$. Panel (**D**) shows the percentage error PE, where the mean value corresponds to the mean percentage error (MPE). In all panels, medians are indicated by annotations, means are depicted as magenta points, and outliers are labeled by scorer ID.

Descriptive patterns across scoring regimes Analysis of the distributional patterns in Figure 16 reveals systematic differences in scoring behavior across experimental conditions. Panel A demonstrates that the AI consistently overestimates arousals relative to ground truth by approximately 10–40%, whereas human-AI teams tend to underestimate by about 20–50%. This contrast reflects a fundamental difference in approach: human scorers adopt a conservative strategy, while the AI model is optimized for recall. The provision of explanations (WB) reduces inter-participant variability, with this stabilizing effect most pronounced in QC conditions.

Performance hierarchy across atomic regimes Figure 16B demonstrates that the QC-WB regime yields the highest median accuracy among all conditions, with teams achieving approximately

86% accuracy compared to 80% for the QC-BB and 74% for the Start-WB regimes. Panel C indicates that teams operating in the QC-WB regime achieve the greatest error reduction relative to the solo AI (mean ratio of 2.45), while teams in Start-based regimes perform worse than the AI alone (ratios below 1.0). Panel D reveals that teams in QC regimes experience substantially reduced under-counting bias compared to those utilizing AI assistance from the start, with QC-WB showing the least negative bias (-14%) compared to Start-BB (-43%).

Scope of inference and design Following the statistical modeling approach described in Section 2.6.2, we analyze the log-transformed AI-Baseline Improvement Ratio $y_{\mathrm{RGT}} = \log R_{\mathrm{GT}}$ and percentage error PE. Within QC we additionally compare team to human solo via paired t-tests on A_{GT} using the methodology outlined in Section 2.6.4.

All inferential results are restricted to the four atomic regimes (Start/QC × WB/BB), following the experimental design outlined in Section 2.6.1.

Statistical Analysis of AI-Baseline Improvement Ratio The omnibus analyses in Table 12 reveal a significant main effect of Timing, accompanied by a large effect size (η_{p}^2=0.66), while neither the AI main effect nor the interaction reach statistical significance.

The planned contrasts in Table 13 confirm that, on average, teams reduce the error rate in the *QC* condition relative to *Start* by approximately a factor of four (ratio = 0.26, 95% CI: [0.15, 0.53]), as indicated by the Timing main contrast. This substantial improvement aligns with the visual patterns observed in Figure 16C, where QC regimes consistently show ratios above 1.0 (indicating team improvement over AI) while Start regimes fall below 1.0. The statistical methodology is described in Section 2.6.3.

Statistical Analysis of Percentage Error PE The statistical results in Tables 14 and 15 indicate a significant effect of Timing (Start vs. QC) on percentage error PE, which quantifies systematic over- or under-counting relative to ground truth.

The RM-ANOVA shows a highly significant main effect of Timing ($p = 0.0044$), indicating a large effect size ($\eta_{\mathrm{p}}^2 = 0.71$). This is supported by the planned-contrast analysis, where the Timing main contrast yields a Holm-adjusted permutation p-value of 0.023 and a substantial effect estimate of -0.24 (95% CI: [-0.34, -0.14]). This means that, on average across AI transparency levels, QC regimes reduce under-counting by about 24 percentage points on the B-scale, relative to Start (absolute shift towards zero), consistent with the visual patterns in Figure 16D. In contrast, neither the AI transparency main effect (WB vs. BB) nor the Timing × AI interaction reach statistical significance in either the omnibus or planned-contrast tests at this sample size.

Comparative Analysis: Teams versus Human Solo Performance Qualitative inspection of raw annotated arousal counts (Figure 16A) reveals that, in both QC regimes, the median arousal count for teams is closer to the ground truth than that of human solo scorers. Notably, in the QC-WB condition, the distribution of team counts is substantially tighter than for solo scoring, with reduced inter-participant variability.

Quantitatively, the coefficient of variation (CV) in QC-WB decreases from 0.47 for human solo to 0.37 for teams, and the mean absolute error (MAE) relative to the ground truth drops from 35 to 21. In the QC-BB condition, improvements are more modest, with the CV decreasing from 0.53 to 0.51 and the MAE reducing from 31 to 27. These improvements are consistent with the visual patterns in Figure 16B, where QC-WB shows the highest accuracy values.

The results from Table 16 on the *count accuracy* confirm that teams tend to exceed human solo performance in QC, especially with WB (mean difference = 0.098, 95% CI: [-0.042, 0.22]), but the differences do not reach statistical significance at this sample size.

Individual Scorer Heterogeneity and Adaptive Strategies Participant heterogeneity is evident across all panels of Figure 16 and provides valuable insights into regime effectiveness. A par-

Table 12: **RM-ANOVA results for $y_{\mathrm{RGT}} = \log R_{\mathrm{GT}}$.** All effects have $df_{\mathrm{effect}}=1$ and $df_{\mathrm{error}}=7$. Partial eta-squared (η_{p}^2) quantifies effect size. Only the Timing main effect is significant.

Effect	$F(1,7)$	p	η_{p}^2
AI main(WB vs. BB)	2.2	0.18	0.24
Timing main (Start vs. QC)	**14**	**0.0075**	**0.66**
Timing × AI Interaction	2.2	0.18	0.24

Table 13: **Planned-contrast test results for y_{RGT} (back-transformed as ratios) in the 2×2 repeated-measures design.** Each row presents the paired t-test statistic, permutation p-value with familywise error control via Holm correction (applied to the two main effects while the interaction uses raw permutation p), the estimated effect ratios (with bootstrap 95% confidence intervals), and within-subject Cohen's d_z effect size (with bootstrap 95% confidence intervals). The results align with the RM-ANOVA findings, demonstrating a significant main effect of Timing. Specifically, the QC condition improves over the Start condition on average by approximately a factor of four.

Contrast	$t(7)$	p_{perm}	Ratio [95% CI]	d_z [95% CI]
AI (WB vs. BB)	1.5	0.18	1.6 [0.87, 2.8]	0.53 [–0.15, 2.1]
Timing (Start vs. QC)	**–3.7**	**0.039**	**0.26 [0.15, 0.53]**	**–1.3 [–5.0, –0.45]**
Interaction	–1.5	0.25	0.32 [0.079, 1.2]	–0.53 [–1.3, 0.15]

Table 14: **RM-ANOVA for percentage error** PE**.** Each effect has $df_{\mathrm{effect}}=1$ and $df_{\mathrm{error}}=7$. Only the Timing main effect is significant.

Effect	$F(1,7)$	p	η_{p}^2
AI main (WB vs. BB)	0.30	0.60	0.041
Timing main (Start vs. QC)	**17**	**0.0044**	**0.71**
Timing × AI Interaction	0.75	0.41	0.097

Table 15: **Planned-contrast test results for percentage error** PE **(raw scale) in the 2×2 repeated-measures design.** Each row reports the paired t-test statistic, permutation p-value with Holm correction for the two main effects (the interaction uses the raw permutation p), the estimated effect differences with bootstrap-derived 95% confidence intervals, and within-subject Cohen's d_z effect size (also with bootstrap 95% confidence intervals). These results are consistent with the RM-ANOVA, revealing a significant main effect of Timing. On average, the percentage error is significantly lower in the QC condition compared to Start, with a reduction of approximately 24 percentage points.

Contrast	$t(7)$	p_{perm}	Estimate [95% CI]	d_z [95% CI]
AI (WB vs. BB)	0.55	0.60	0.026 [–0.057, 0.12]	0.19 [–0.84, 0.96]
Timing (Start vs. QC)	**–4.1**	**0.023**	**–0.24 [–0.34, –0.14]**	**–1.5 [–2.8, –1.1]**
Interaction	0.87	0.47	0.17 [–0.16, 0.56]	0.31 [–0.56, 0.97]

Table 16: **Team vs. solo on GT accuracy A_{GT} within QC**. Differences are team minus solo. $p_{\mathrm{perm,Holm}}$: Holm-adjusted within the two QC cells.

Regime	n	Team mean	Solo mean	Diff [95% CI]	$t(7)$	$p_{\mathrm{perm,Holm}}$
QC, WB	8	0.82	0.72	0.098 [–0.042, 0.22]	1.4	0.46
QC, BB	8	0.78	0.73	0.046 [–0.040 0.12]	1.0	0.46

ticularly illustrative case is scorer *HU-5*, whose behavior reveals complex patterns of AI interaction and adaptive strategies.

HU-5 consistently appears as an outlier across all panels, exhibiting systematic over-counting behavior. In Panel A (Total Event Count), *HU-5* records the highest arousal counts across most regimes, with especially pronounced values under QC conditions. This over-counting pattern is mirrored in Panel D (Percentage Error), where *HU-5* consistently displays positive values, indicating persistent overestimation relative to ground truth.

Consequently, *HU-5* achieves lower accuracy than other teams (Panel B) and performs worse than the AI alone in several regimes (Panel C). A notable exception occurs in the Start-BB regime, where *HU-5* achieves the highest performance among all teams. This outcome results from an adaptive strategy documented through direct observation: in this final regime, *HU-5* reviewed only AI-suggested events, leaving intervals between suggestions unscored and accepting nearly all AI-suggested events. Given the AI's strong performance in the Start-BB regime, this selective reliance strategy proves particularly effective.

This case demonstrates the dynamic interplay between user trust, strategy selection, and regime design, highlighting how individual differences in AI reliance can substantially influence performance outcomes.

Summary and Clinical Implications The count-based evaluation establishes a clear hierarchy of factors influencing human-AI team performance in arousal scoring. *Timing* emerges as the primary determinant, with QC-based workflows substantially outperforming Start-based approaches across all evaluated metrics. Teams utilizing AI during QC achieve greater error reduction relative to AI alone (R_{GT} ratio of 0.26, $p = 0.032$) and exhibit significantly reduced count bias compared to ground truth ($B = -0.24$, $p = 0.023$). In contrast, *AI transparency* confers only modest, non-significant benefits, particularly for error reduction in QC contexts. The interaction between Timing and AI transparency is likewise non-significant, indicating that the advantage of QC is robust across both WB and BB conditions.

These findings have important implications for clinical implementation. QC-based workflows produce the most favorable distributional patterns. The visual evidence in Figure 16 demonstrates that QC-WB brings teams closer to ground truth than any other regime, supporting a workflow that separates an initial human pass from a quality control step, with preliminary evidence suggesting that transparency may provide additional stabilization without undermining human judgment.

4.5 Time Demand and Efficiency of Arousal Scoring

To understand the practical implications of AI assistance on workflow efficiency, we examined the time demand associated with each phase of the user study. Figure 17 presents boxplots summarizing the duration of all relevant scoring phases across the cohort. The figure shows the median (red line), interquartile range (box), whiskers, and individual data points (blue circles), with outliers highlighted in red. Annotations mark the medians for each distribution.

Phases 1 and 8, covering the introductory video (approx. 13 minutes) and the closing questionnaire, are excluded from the plot. The final column, "Total Study Duration," captures the cumulative time per participant, including these omitted phases as well as breaks and technical delays. While the overall study lasted around 5.5 hours per participant, individual differences are considerable, likely reflecting differences in user affinity with digital tools and adaptive behavior under fatigue.

The results reveal two prominent patterns: First, across both task settings (quality control (QC) and assistance from the start) white box (WB) AI support required notably more time than black box (BB) support. Specifically, the median durations of the WB phases (66 and 36 minutes) are approximately twice those of their BB counterparts (38 and 18 minutes, respectively). This increased time demand is consistent with the cognitive overhead of interpreting explanations and the study protocol's requirement that participants meaningfully engage with these explanations. Our observations indicated that participants increased their speed after becoming familiar with the initial explanations. Notably, the WB durations also show greater dispersion, indicating individual

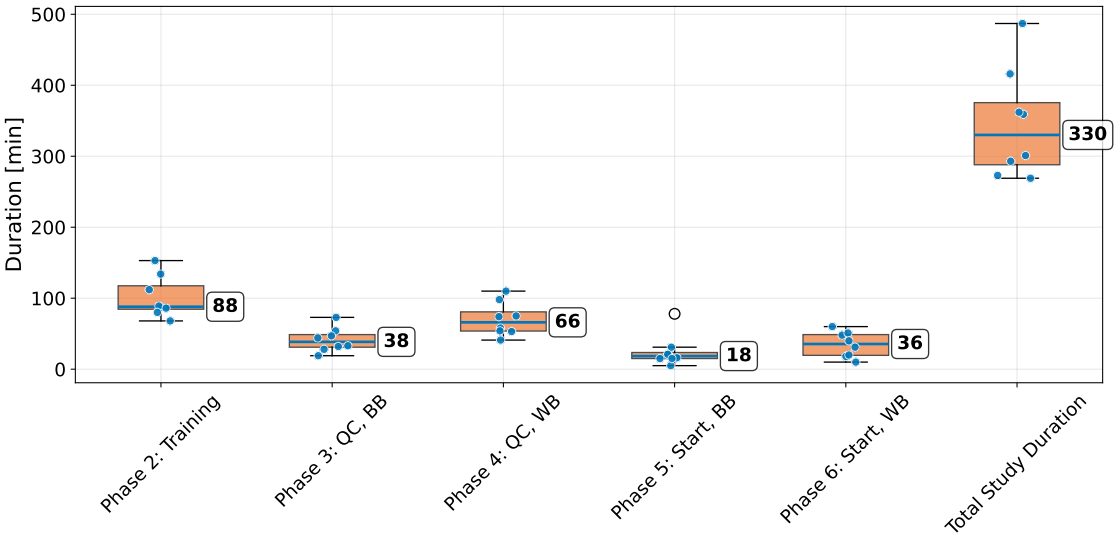


Figure 17: **Time Demand** across user study phases. The boxplots illustrate the distribution of durations for each phase, with inlier points represented by blue circles and outliers by empty circles. The labels denote the median duration in minutes for each phase. Notably, AI support provided during the quality control (QC) phase results in scoring sessions that are approximately twice as long compared to when AI support is given right from the start.

variation in speed. Some participants experienced similar time demands for both WB and BB support, indicating that efficiency can be achieved with increased familiarity.

Second, we observe that QC phases are generally more time-consuming than start-support phases, with QC sessions taking almost twice as long regardless of AI explainability. This makes sense given the sequential workflow: participants had to reassess the data in full, rather than make decisions immediately based on AI input. The increased duration thus stems from re-scanning the timeline, either by revisiting specific events flagged by the AI or conducting a comprehensive second pass through the data. Importantly, this QC overhead represents a conservative upper bound for time demand. In real-world deployments, practice and interface optimization would likely reduce this burden significantly.

One particularly revealing case merits discussion: a participant in the final phase (*Start, BB*) completed the task in just 5 minutes, the shortest session across all participants and conditions. This user had developed enough trust in the AI to only evaluate AI-suggested events, effectively skipping segments without flagged arousals. This behavior illustrates the impact of AI design on user trust and reliance: the system, which prioritizes recall over precision, may have led users to over-trust its coverage. While encouraging from a workflow efficiency standpoint, this behavior underscores the importance of monitoring how AI shapes user attention and vigilance.

In summary, while white box AI introduces an initial time cost, its efficiency may converge toward black box levels as users gain fluency. Similarly, although QC demands more time than direct assistance, it may offer benefits in user autonomy and error-checking that justify this investment, especially during onboarding. These dynamics are crucial for deployment planning in time-sensitive environments such as clinical diagnostics or real-time monitoring.

4.6 Evaluation of Questionnaires

This section primarily delves into the secondary objectives *SO1-SO9*, while also offering valuable insights into the primary objective *PO3* (refer to Section 3.6.1). The numerical findings from the questionnaires are illustrated in Figures 18, 19, and 20. Free-text excerpts have been slightly edited for fluency, with the original responses available in Appendix E. All questions and responses have

been translated from German to English.

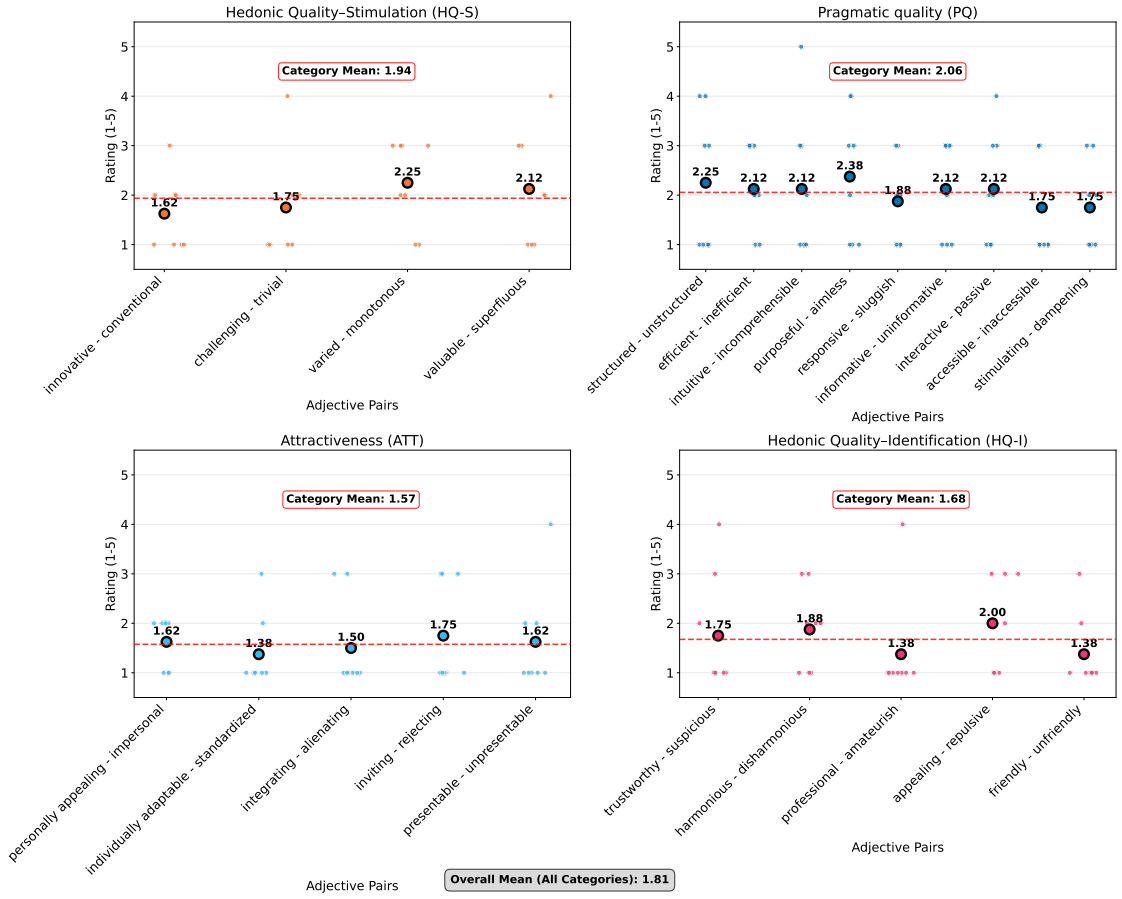


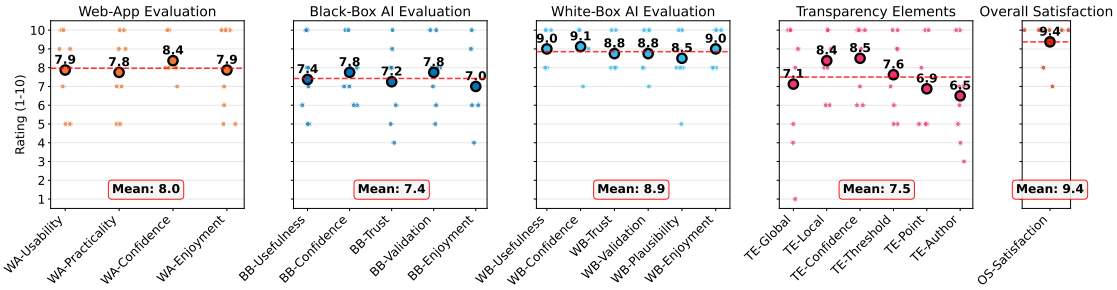
Figure 18: **Evaluation of User Experience** utilizing the AttrakDiff questionnaire. This figure presents participant ratings across four dimensions: Hedonic Quality-Stimulation (HQ-S), Pragmatic Quality (PQ), Attractiveness (ATT), and Hedonic Quality-Identification (HQ-I). It emphasizes both the category means and the overall mean, with ratings on a 1-5 scale, where lower values indicate more favorable assessments. Participants rated all dimensions with a mean score of 2 or lower.

4.6.1 Usability and User Experience of the Web App

Before delving into the AI support, we first examine the usability and user experience of the web app. It is important to note that the AttrakDiff ratings are on a scale from 1 to 5, where lower scores are preferable, while the rating questions use a scale from 1 to 10, where higher scores are more favorable. The web app's usability is commendable, as evidenced by the Pragmatic Quality ratings in the AttrakDiff questionnaire (Figure 18), which show an average score of 2.06. Additionally, participants awarded the web app an explicit usability rating of 7.9 and a practicality score of 7.8 (Figure 19).

The user experience is also highly rated, as reflected in the excellent hedonic quality ratings, with an average stimulation score of 1.94 and an identification score of 1.68 on AttrakDiff, alongside an average attractiveness rating of 1.57 (Figure 18). These findings align with the explicit ratings for web app confidence at 8.4 and web app enjoyment at 7.9 (Figure 19).

The positive user experience of the web app is further supported by participant feedback in the form of free-text responses. One participant remarked that the web interface felt *very detailed and*



For all questions, participants were asked: "Rate the following aspects on a scale from 1 (lowest rating) to 10 (highest rating)"

Web-App Evaluation:
- WA-Usability: How do you rate the user-friendliness of the web app?
- WA-Practicality: How do you rate the practical usability of the web app?
- WA-Confidence: How confident did you feel during manual evaluation (without AI support) of arousals?
- WA-Enjoyment: How much fun do you have with manual arousal evaluation using the provided application?

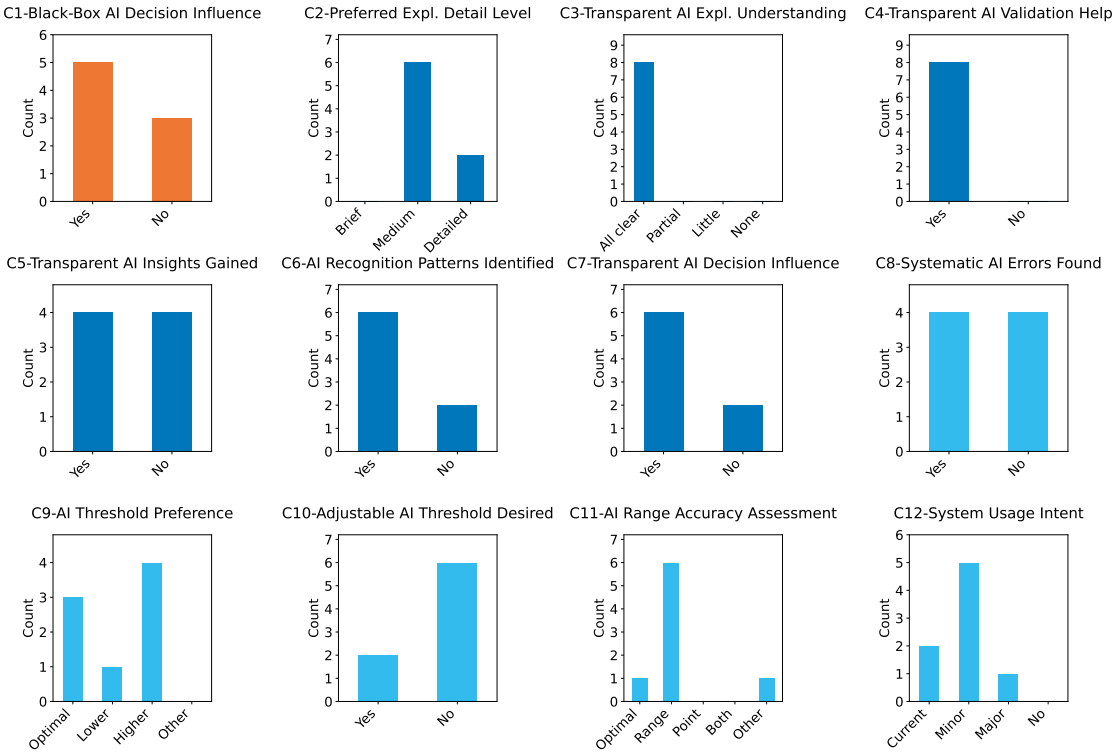
Black-Box AI Evaluation:
- BB-Usefulness: How useful do you find the support from the Black-Box AI overall?
- BB-Confidence: How confident did you feel during arousal evaluation with Black-Box AI support?
- BB-Trust: How much do you trust the results of the Black-Box AI?
- BB-Validation: How easy was it for you to validate the arousals determined by the Black-Box AI?
- BB-Enjoyment: How much fun do you have with arousal evaluation using the Black-Box AI?

White-Box AI Evaluation:
- WB-Usefulness: How useful do you find the support from the transparent AI?
- WB-Confidence: How confident did you feel during arousal evaluation with transparent AI support?
- WB-Trust: How much do you trust the results of the transparent AI?
- WB-Validation: How easy was it for you to validate the arousals determined by the transparent AI?
- WB-Plausibility: How plausible are the AI explanations for arousal determination?
- WB-Enjoyment: How much fun do you have with arousal evaluation using the transparent AI?

Transparency Elements:
- TE-Global: How helpful is the global dataset explanation (bar chart "Global relevance ...") for creating transparency of the AI decision?
- TE-Local: How helpful is the local explanation for an arousal (graph in footer "Relevance of data points ...") for creating transparency of the AI decision?
- TE-Confidence: How helpful is the probability channel (Confidence Score) for the confidence of the AI model for creating transparency of the AI decision?
- TE-Threshold: How helpful is the indication of the probability threshold in the probability channel (horizontal dashed line) for creating transparency of the AI decision?
- TE-Point: How helpful is the indication of the most likely point for an arousal start (vertical dashed line) for creating transparency of the AI decision?
- TE-Author: How helpful is the indication of author confidence (bar with green gradient) in the event details for creating transparency of the AI decision?

Overall Satisfaction:
- OS-Satisfaction: How satisfied are you overall with participating in the study?

Figure 19: **Analysis of Rating Questions** derived from participant feedback. This figure presents evaluations for the Web-App, Black Box, and White Box AI, as well as Transparency Elements and Overall Satisfaction, with the mean values indicated by a red line for each category. Ratings are on a 1-10 scale, where higher values denote greater favorability. Notably, ratings for the white box AI approach a score of 9, demonstrating a preference for transparency and explanatory features, as they are over 40% closer to the ideal score of 10 compared to the black box AI ratings.



C1: Has the support from the Black-Box AI influenced your decisions? If yes, please give examples.
C2: Which of the provided detail levels of explanations for AI-determined arousals is most useful for you?
Response Options: Brief: Brief Detail | Medium: Medium Detail
C3: How well could you understand the explanations of the transparent AI? Were there aspects that were unclear?
Response Options: All clear: Everything understandable | Partial: Partially understandable | Little: Little understandable | None: Not understandable
C4: Did the explanations for the AI-determined arousals help you with their validation?
C5: Did you gain interesting or unexpected insights from the transparent AI explanations about the AI's approach to arousal determination? If yes: What are they?
C6: Could you identify one or more recurring patterns by which the AI recognizes arousals? If yes: What are they?
C7: Has the support from the transparent AI influenced your decisions? If yes, please give examples.
C8: Did you find systematic errors that the AI model makes? If yes: Which ones?
C9: The AI model was optimized to miss as few arousals as possible. For this reason, there are more frequently false positive determined arousals, i.e., the model marks arousals more frequently where there actually are none. For this purpose, the threshold for detecting an arousal was set relatively low. How do you find this approach?
Response Options: Optimal: Exactly right | Lower: An even lower threshold would be better (fewer false negatives, more false positives) | Higher: A higher threshold would be better (fewer false positives, more false negatives)
C10: Would it be helpful if you could adjust the decision threshold of the AI model to set at which confidence level the AI marks an arousal event?
C11: When the AI had correctly recognized an arousal, how good was in your opinion generally the agreement of the AI-determined range in which the arousal start likely lies with the actual start of the arousal? Was the range too large or the most likely point (vertical dashed line) too far from the actual start point? Please consider in your answer what degree of agreement is required for the practical usability of the AI support for arousal evaluation.
Response Options: Optimal: Exactly right | Range: AI-determined range was too large | Point: Most likely point too far away from actual point | Both: Range too large and most likely point too far away
C12: Could you imagine using the present system for arousal evaluation in the sleep laboratory?
Response Options: Current: Yes, in current form | Minor: Yes, with minor improvements | Major: Yes, with major improvements

Figure 20: **Analysis of Categorical Responses** derived from participant feedback. This figure depicts the distribution of categorical responses to various inquiries concerning black box (orange) and white box (dark blue) AI support, as well as different elements related to AI and the web application (light blue). Remarkably, all participants reported a complete understanding of the provided explanations, and unanimously found the transparent AI support beneficial for validating AI-suggested arousal events. Additionally, seven out of eight participants expressed a willingness to utilize the AI-powered Decision Support System (DSS) with minor modifications or in its current state, reflecting a high degree of acceptance and usability.

professionally created once one gets an overview, while another simply stated: *I find the web app user-friendly.*

In conclusion, the web app is perceived as ranging from good to very good by participants, which is a crucial prerequisite. This perception suggests that AI assistance is not hindered by poor overall usability and user experience of the web app.

4.6.2 PO3 - AI Assistance Timing

In the free-text responses, participants were asked about their preference for starting with AI support or beginning with manual scoring followed by AI-based quality control. Six out of eight participants expressed a clear preference for initiating with AI support, with one participant highlighting *increased confidence and speed* as the rationale. Another participant offered a more nuanced perspective, noting that *AI support from the start is certainly easier - for experienced evaluators! (somewhat tempting for newcomers).* This insight points to a potential risk for newcomers, who may become overly reliant on AI suggestions or find it challenging to critically evaluate them. While this study does not explore this aspect in depth, it remains a crucial consideration for introducing AI assistance in real-world applications.

4.6.3 PO3 - AI Assistance Usefulness of Transparency

Participants rated the usefulness of AI support on a $1-10$ scale, with black box support receiving a score of 7.4 and transparent (white box) support achieving a higher score of 9.0 (Fig. 19). This preference for white box AI is echoed in free-text responses, where several participants expressed a clear preference for it, stating comments like *White Box was better for me* or *Transparent White Box AI support was my favorite*, while no such endorsements were made for black box AI.

Furthermore, a significant majority reported that AI, particularly the transparent variant, influenced their decisions (75% for white box compared to 62.5% for black box). When asked for examples to illustrate this influence, responses regarding black box AI assistance were mostly generic (e.g., *Regarding confidence* or *one or the other arousal*). In contrast, feedback on white box AI assistance included both general comments (e.g., *occasionally*, or *positively influenced*) and specific examples (e.g., *PLM seen and I didn't but the AI was right.* or *Breathing pattern /PLM*, where PLM stands for Periodic Leg Movement), indicating a deeper level of user engagement and insights gained with white box AI. These aspects are further discussed in the following sections.

4.6.4 SO1 - Trust in AI Predictions

Trust scores mirror the utility pattern. Manual scoring confidence was already high (mean = 8.4), but black box trust dropped slightly to 7.3. Transparent AI, by contrast, achieved a mean trust rating of 8.8 meaning that participants overall expressed slightly more trust in the transparent AI than in their own manual scoring. The AttrakDiff adjective pair "trustworthy – suspicious" gives the application a rating of 1.75 (on a 1–5 scale where lower values are positive) leaning strongly towards the first adjective, suggesting that the interface as a whole felt trustworthy (Fig. 18). Taken together, the data indicate that explanations bolster user trust even beyond the baseline confidence derived from domain expertise. Several free-text notes connect confidence and trust directly to the presence of explanations, with a spectrum of reliance and trust from high to low:

- *95% correctly marked*
- *The explanation partially convinced me to score an arousal*
- *Explanations were the confirmation for me*
- *In case of uncertainty explanations definitely helped and were often conclusive*
- *Explanations helped for arousals where I wasn't quite sure*
- *I thought about why the blackbox might have marked this now and with the whitebox I looked at why it marked it and then as always decided who was right*

- *The explanations helped but were not decisive in my decisions*

We wish to emphasize that the statement *95% correctly marked* originates from participant *HU-5*, who is identified as an outlier both at the event-based level (refer to Section 4.2.4) and in terms of total event count (refer to Section 4.4). Both analyses also reflect sub-par performance for this participant. The combination of these quantitative results and the statement strongly indicates misplaced trust.

4.6.5 SO2 - Perceived Comfort

Self-reported confidence while working with the AI aligns with the trust pattern: 7.8 for black box versus 9.1 for white box assistance. Overall study satisfaction was exceptionally high (mean = 9.4), implying that the experimental workload and tooling were acceptable. Thus, transparent assistance not only inspired trust but may also have fostered a comfortable working atmosphere. This is backed by the good to very good ratings on Usability and User Experience of the Web App (Section 4.6.1).

4.6.6 SO3 - Ease of Validation

Participants found black box predictions reasonably easy to validate (mean = 7.8), yet transparent predictions were judged even easier (mean = 8.8). All participants answered "yes" when asked whether explanations helped them in validation (C4 in Fig. 20), underscoring the perception that explanations accelerated or simplified verification. This tendency is backed by the response of one participant, stating *With the White Box AI, the evaluation went much faster because you can rely very much on the AI having marked everything important, unlike manual evaluation.* The participant was actually 1.5 times faster with WB AI support compared to BB AI support when assistance was provided from the beginning. However, during the QC phase, the participant was approximately twice as slow. Since the analysis in Section 4.5 shows that overall, phases with WB support took on average about twice as long as BB support phases, the experience of faster validation with WB AI seems to be subjective than backed by objective results.

Beyond the textual indicators of the value of explanations for validation discussed in *S2* (Section 4.6.4), a participant noted that *explanations support and you take a closer look at it again*, and another participant who prefered transparent AI stated: *One actually gets influenced and questions one's decisions again*, both indicating that explanations encourage a more thorough validation.

A participant highlighted a notable benefit of the local explanations for validation: *Due to lack of space, additional channels could simply be viewed in the explanations, especially when the AI has already classified them as relevant, they are very interesting.* This observation emphasizes the utility of local relevance plots, which prioritize channels from most to least relevant, thereby reinforcing the overall value of the explanations.

4.6.7 SO4 - Plausibility of Explanations

Plausibility ratings for white box explanations averaged 8.5. Every participant declared the explanations completely understandable (C3 in Fig. 20); no one selected "partially" or "little" understandable. Within individual transparency elements (see Fig. 19), local explanations (mean = 8.4) and the confidence channel (mean = 8.5) were deemed most helpful, whereas global relevance statistics, the most-likely event start indicator, and aggregate-confidence scores were less helpful (means = 7.1, 6.9, and 6.5, respectively). This pattern suggests that detailed, event-specific visuals resonate more strongly than aggregate information. Concerning the transparency elements, one participant stated *I missed the threshold/confidence* when being left with black box AI support. In the free-text answers on how to improve the transparent AI support, one participant suggested that the AI model should prioritize EEG frequency increases. This would be in accordance with American Academy of Sleep Medicine guidelines, wherease the current modelling approach is guideline-agnostic. Another requested *shorter markings*, hinting at interval-length issues discussed in *SO9* (Section 4.6.12).

4.6.8 SO5 - Satisfaction and Enjoyment

Enjoyment ratings again favour transparency: 7.0 under black box versus 9.0 under white box conditions. When combined with the high overall-satisfaction score, these findings imply that explainability may enhance intrinsic motivation, a non-trivial benefit given the repetitive nature of arousal scoring.

One scorer expressed that the AI system *marked 95% of the arousals correctly*, orally stating *No other sleep system can match this*, capturing the enthusiasm behind the numbers.

4.6.9 SO6 - Preferred Level of Detail

Six of eight participants preferred a *medium* level of explanatory detail, while the remaining two opted for *detailed*. Nobody chose "brief," indicating that terse justifications are insufficient, and very granular output is favored by some users. This suggests that concise but information-rich visuals strike the optimal balance, where the slight tendency towards the more detailed level of detail is consistent with the findings of *SO4* (4.6.7). No free text reported confusion about over-complexity.

4.6.10 SO7 - Insights into AI Reasoning

Half of the cohort reported gaining novel insights from the explanations, and 75% could articulate recurring patterns with respect to various modalities that trigger arousal predictions: *Pulse rate*, *PLMs (don't always lead to an arousal) and desaturations with breathing events*, *Heart rate increase*, *Thorax and abdomen breathing, EMG*, *EEG*, *Breathing pattern /PLM*. Such meta-knowledge can foster calibrated trust, and may gradually shift human expertise toward more consistent scoring standards. The free text responses to the question about insights gained from explanations about transparent AI arousal determination methods indicate that participants observed the AI's comprehensive use of all available data and noted a significant divergence in the data focus compared to human evaluators: *The AI really uses all available data*, *The data that the AI uses for assessment differs greatly from the data that an evaluator normally pays attention to.* One participant expressed surprise that the AI marked arousals when the patient was awake, indicating potential areas for enhancing AI alignment with current scoring standards.

4.6.11 SO8 - Decision-Threshold Preference

Only three participants considered the low decision threshold appropriate; four wanted it higher, one lower. Yet six of eight opposed a user-adjustable slider (C9 and C10 in Fig. 20), suggesting that participants prefer a little less conservative default but are reluctant to fine-tune model sensitivity themselves. Future iterations could therefore expose two or three vetted presets rather than a free slider.

Free-text answers reveal that the reasons of those who favored a higher threshold are related to time-efficiency of the workflow: *Too many false arousals. I get somewhat delayed by this*, *I score an arousal rather quickly, rather than having to sight and then possibly decide*, and *I'd rather draw a few "forgotten" arousals myself than having to delete many again.*.

4.6.12 SO9 - Temporal Accuracy of Predictions

Six of eight participants felt the predicted arousal range was too wide, while merely one person considered it *exactly right*. No respondent complained that the most-likely point was systematically misplaced, pointing to interval width, not point accuracy, as the main usability bottleneck.

Those, wishing for a smaller range, expressed either a size of $\pm$*3 seconds* or a factor of *0.3 to 0.5* of the current range. Two participants clearly opted for only displaying the most likely point stating *the line was exactly right*, and that an improvement would be a *limitation of the marked area to the blue line.*

An obvious improvement would therefore be to only display the most likely point, and to keep the whole range through the confidence channel which was highly appreciated (see *SO4*, Section 4.6.7).

4.6.13 Conclusion of the Questionnaire Study

The questionnaire study yields three overarching take-aways that extend and corroborate the quantitative performance results.

1. The platform is ready for AI integration. Participants rated both pragmatic and hedonic qualities of the web application as *good to very good* (Section 4.6.1), removing basic usability as a confounder when interpreting the effects of algorithmic support.

2. Transparency is the dominant driver of acceptance. Across all key dimensions – perceived usefulness and trust (Sections 4.6.3, 4.6.4), comfort (Section 4.6.5), validation ease (Section 4.6.6), explanatory plausibility (Section 4.6.7), and enjoyment (Section 4.6.8) – the white box assistant outperformed the black box by one to two Likert points. Free-text notes confirm that explanations not only foster trust but also stimulate critical reflection rather than blind acceptance.

3. Clinically relevant refinements are still required. Users welcome AI support from the start (Section 4.6.2), yet half of them request a *higher* decision threshold and three-quarters deem the current onset window too wide (Section 4.6.11, Section 4.6.12). They also ask for stronger emphasis on EEG features and prevent the AI from scoring arousals when the patient is awake to align the model with American Academy of Sleep Medicine guidelines (Sections 4.6.7 and 4.6.10). These points can be addressed with an optimization closer to the F1 than the F2 score, a narrower or even point-based visual range focused on the most-likely start, having one EEG channel always prominently visible in all explanations, and (in case of events during wakefulness) more extensive post-processing of the AI's predictions.

Practical implication. A transparent, event-centred explanation layer is not an optional add-on but a prerequisite for clinical adoption. With minor interface and model tweaks, the system could move from promising prototype to a trusted co-scorer in routine sleep-laboratory workflows. The majority of participants concur with this evaluation, as reflected in the results C12 in Fig. 20. Specifically, two participants are ready to adopt the system in its current state, five participants suggest minor enhancements, and only one participant advocates for significant modifications.

5 Discussion

Our study addresses three fundamental research questions about human-AI collaboration in clinical arousal scoring (see Section 1), which we summarize as follows:

RQ1: Is AI assistance preferable to unaided human scoring?

RQ2: Does transparent (white-box, WB) assistance outperform opaque (black-box, BB) assistance?

RQ3: Does it matter whether the AI is consulted from the start or only in a post-hoc quality-control (QC) pass?

The evaluation of the effectiveness of AI assistance depends critically on the reference standard and evaluation measures employed. Our comprehensive analysis integrates event-level performance, count-based clinical measures, temporal efficiency, and user experience to provide insights into human-AI collaboration that would be missed by any single evaluation approach (Sections 4.2 – 4.6). This multi-faceted evaluation is particularly valuable for clinical decision support systems, where both technical performance and user acceptance are critical for successful implementation.

5.1 The Critical Role of Evaluation Standards

Our dual-ground-truth analysis shows that conclusions about AI assistance effectiveness, effectiveness, addressing **RQ1**, are contingent on the evaluation standard. Under the relatively neutral *consensus* reference, the analyses (Section 4.2.2) reveal substantial inter-rater heterogeneity and clustering structure, indicating that the consensus aggregate is not closely aligned with the *CPS* standard used to train the AI, even though that standard entered the pool via participant HU-9.

In this setting, AI assistance does not confer measurable gains when judged against the consensus benchmark (Table 5).

By contrast, when evaluated against the *CPS* ground truth, both AI and human-AI teams significantly outperform human solo scoring (Table 7). Concretely, the distribution of human solo F1 scores under CPS has mean μ=0.42 and standard deviation σ=0.09, whereas human-AI teams improve to μ=0.56 with reduced variability σ=0.05 (Figure 15; Section 4.3.1). The pattern is consistent with AI assistance shifting human scoring toward the standard underlying model training. Given the typically poor inter-rater reliability in arousal scoring (in a recent study, a multi-center intraclass correlation of ICC=0.41 was reported Pitkänen et al., 2024, indicating "poor" inter-rater reliability), such a shift is not unexpected and, importantly, is accompanied by a clinically desired reduction in variability.

Taken together, for **RQ1**, AI assistance for arousal detection is advantageous when the objective is alignment with a clearly specified clinical standard (here, CPS), whereas it offers no benefit under a neutral but heterogeneous consensus reference. These findings place heightened importance on the careful curation and governance of the training standard for AI intended for clinical decision support. Moreover, our evidence of multiple coexisting scoring tendencies among experts (Section 4.2.2) suggests that consolidating and harmonizing standards should precede broad deployment.

The count-based analysis adds the perspective of clinical decision-making (Section 4.4; Figure 16). We observe systematic biases: the AI *overestimates* arousals by roughly 10–40% while humans *underestimate* by 20–50%. The outcome for the AI is expected since it was optimized for the F2 score. Human-AI teams reduce, but do not eliminate, this gap and magnitude of the reduction is regime-dependent. In the most favorable regime, transparent AI assistance as quality control, both variability and absolute error show a downward trend, as the coefficient of variation decreases from 0.47 (human solo) to 0.35 (team), and MAE drops from 34.5 to 20.8 events (Section 4.4). Also, the median team discrepancy falls below 15% of the ground-truth total (cf. Section 4.4). While this pattern in the count-based evaluation mirrors the improvements observed in the event-level analysis, it does not reach statistical significance (Table 16).

5.2 Impact of Transparency on Performance and User Experience

Findings on event-level performance (Section 4.3.2) directly address **RQ2**: transparent assistance consistently outperforms opaque assistance. Statistically, white-box (WB) assistance yields an *average* 18% improvement in relative F1 over black-box (BB) assistance with a significant AI transparency×Timing interaction (Table 10). The transparency benefit is most pronounced in quality control (QC), where WB exceeds BB by about 30% on average (Table 11).

Count-based evaluation shows that transparency provides modest, although non-significant additional benefits. Specifically, the QC-WB teams achieved the highest accuracy and lowest bias for underestimation of arousal counts, but the advantage of transparency over black-box assistance did not reach statistical significance at this sample size. Nonetheless, visual patterns suggest that transparency may help stabilize team performance and reduce inter-participant variability, particularly in QC workflows.

User-experience results align with these effects. Participants rate WB 1–2 Likert points higher across usefulness, confidence, trust, ease of validation, and enjoyment (Section 4.6). On the 1–10 usefulness scale, WB narrows the gap to the ideal score of 10 by *more than 40%* relative to BB (Figure 19; Section 4.6.3). Together, these findings indicate that transparency enhances both performance and acceptance, key prerequisites for clinical adoption.

5.3 Timing Effects and Workflow Strategy

Our findings provide nuanced answers to **RQ3**. For event-level measures, although timing is not a significant main effect (Table 10), the significant AI×Timing interaction indicates that transparency matters most at quality control (QC) timing (Table 11).

For clinically oriented count-based measures, timing is the dominant factor (Section 4.4): Compared to Start timing, QC regimes improve the AI-baseline Improvement Ratio by roughly a factor

of four on average (the timing ratio is about 0.26, see Table 13) and substantially reduce systematic under-counting (the timing effect on the percentage error PE is reduced by approximately 24 percentage points, see Table 15).

Survey feedback on the preferred timing of AI assistance (see Section 4.6.2) indicates a clear inclination among participants to utilize AI support from the outset of scoring, rather than reserving it for a subsequent quality control step. This early integration was seen as beneficial for workflow efficiency and user confidence, particularly among experienced scorers. Notably, one participant highlighted the need for caution, as immediate AI involvement could potentially foster over-reliance or diminish critical assessment, especially for less experienced users.

Taken together, these findings highlight a key tension: while objective performance metrics point to the greatest accuracy gains when AI assistance is used as a quality-control step, many users express a preference for integrating AI support from the outset to enhance workflow efficiency and confidence. This underscores the challenge of designing systems that not only optimize measurable outcomes but also align with user preferences and promote sustained engagement.

5.4 Clinical Implications, Time Efficiency, and Workflow Integration

Combining transparency with a QC timing yields the most favorable outcomes across measures. In QC-WB, teams achieve the highest count accuracy (Figure 16B), the strongest improvement over the AI baseline (Figure 16C), and the smallest residual bias (Figure 16D). This regime also shows reduced inter-participant variability (Section 4.4).

These gains however come with time costs. Median duration approximately doubles for WB versus BB and, independently, QC sessions are roughly twice as long as Start sessions (Figure 17; Section 4.5). Some participants, however, already achieved WB speeds comparable to the fastest BB (Section 4.5), suggesting learning effects may mitigate transparency overhead. Furthermore, one participant explicitly expressed a preference for white-box over black-box due to enhanced speed, a preference substantiated by the data for this participant. By contrast, QC's extra pass is inherently costly and unlikely to diminish substantially with experience.

White-box vs. black-box. Given the large performance and acceptance advantages of WB over BB (Sections 4.3.2, 4.6), we judge the additional time to be justifiable. Several participants also appeared to consult explanations selectively. With familiarity, brief verification via explanations may be even faster than speculating about a black-box suggestion.

Quality control vs. start-based. Six of eight participants preferred starting with AI (Section 4.6.2), citing confidence and speed, but QC delivered the most accurate counts, the largest error reduction relative to AI, and the smallest bias (Section 4.4). A practical compromise is therefore context-dependent: For routine scoring, a faster yet less accurate support from the start may suffice. Conversely, for auditing, educational purposes, or complex cases, the more time-consuming quality control approach may prove beneficial.

5.5 Human-AI Collaboration: Beyond Simple Automation

The benefit-ratio analysis indicates teams adopted about 46% of the AI's alignment signal on the median, with wide individual spread (0.10–0.78; Figure 15, Section 4.3.1). While this may indicate a level of critical evaluation which is beneficial, it may also include missed opportunities for improvement. Notably, Sections 4.2.4 and 4.4 document pitfalls of strong alignment and over-reliance (e.g., scorer HU-5), including acceptance of false positives and insufficient correction. Despite such cases, and the fact that teams typically did not *exceed* the solo AI at the event level (Table 8), there are strong reasons to retain humans in the loop: QC-WB teams frequently surpass the AI on clinically salient count-based outcomes (Figure 16C), humans cope better with out-of-distribution cases, and human oversight supports accountability, bias mitigation, and iterative model improvement.

In summary, transparent AI assistance, particularly when applied as a quality-control step, empowers human experts to harness the strengths of artificial intelligence while actively mitigating its

errors. This collaborative approach produces results that are not only more consistent than those achieved by unaided humans, but also more reliable and trustworthy than those generated by AI alone.

5.6 Limitations

Several limitations warrant consideration when interpreting our findings.

The study's reliance on eight participants, while adequate for detecting large effects, limits the ability to identify more subtle interactions and may introduce individual biases that affect generalizability. The small sample size also constrains our ability to detect smaller but potentially meaningful effects, particularly in the interaction analyses.

The dual-ground-truth approach, while methodologically rigorous, introduces its own limitations. The consensus labels were generated by the same annotators involved in the study, potentially introducing residual bias. The CPS ground truth, while representing established clinical practice, reflects individual scoring styles that may not generalize across institutions or time periods. Future research should consider involving independent expert panels to enhance ground truth validity.

The AI model's optimization for high recall (F2 score) may have influenced human adoption patterns and interaction behaviors. A model tuned for precision might yield different collaboration dynamics, suggesting the need for further exploration of alternative optimization strategies and their impact on human-AI interaction.

The study utilized a single explanation type (causal explanations in the form of local relevance plots together with confidence traces), limiting our understanding of how different explanation modalities might affect collaboration. Investigating other types of explanations, such as contrastive or counterfactual explanations, could provide additional insights into optimal explanation design for clinical applications.

Although the study design incorporates counterbalancing and different subjects per phase to mitigate learning effects, some learning curve may still be present. The limited exposure duration (primarily first-time usage) means that long-term effects on user trust, fatigue, and adaptation remain unexplored.

Finally, the user interface design significantly influences collaboration effectiveness, and our specific implementation choices may have affected the results. The iterative design process incorporated expert feedback, but alternative interface designs could yield different interaction patterns and performance outcomes.

5.7 Conclusion

The central lesson of this study is that human-AI collaboration in arousal detection significantly enhances team alignment with the reference standard used to train the AI: teams become more accurate and more consistent, and the key lever is *transparency*. Explanations transform the AI from a mere source of suggestions into a provider of actionable evidence, yielding the most reliable results when transparent support is applied as a targeted *quality-control* step.

Beyond quantitative performance, clinicians express a strong willingness to adopt such systems. Transparent assistance receives higher ratings for usefulness, trust, ease of validation, and enjoyment, with most participants indicating they would use the system with at most minor modifications. However, these benefits come at the cost of increased time, as both transparency and quality control introduce workflow overhead. This underscores the need for *configurable workflows*: transparent AI assistance for quality support should be employed when accuracy is critical, while start-time assistance can address throughput demands, provided safeguards are in place to prevent over-reliance.

Collectively, these findings position human-AI arousal scoring as a *design* challenge as much as a modeling one. Systems that integrate clinically meaningful explanations, flexible timing, calibrated decision thresholds, and precise onset visualization can facilitate alignment with the chosen standard while maintaining human agency. In this way, transparent AI functions as a reliable co-scorer, supporting rather than supplanting expert judgment.

References

Kasun Amarasinghe, Kit T Rodolfa, Hemank Lamba, and Rayid Ghani. Explainable machine learning for public policy: Use cases, gaps, and research directions. *Data & Policy*, 5:e5, 2023.

Kasun Amarasinghe, Kit T Rodolfa, Sérgio Jesus, Valerie Chen, Vladimir Balayan, Pedro Saleiro, Pedro Bizarro, Ameet Talwalkar, and Rayid Ghani. On the importance of application-grounded experimental design for evaluating explainable ml methods. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 20921–20929, 2024.

Afsoon Badiei, Saeed Meshgini, and Khosro Rezaee. A novel approach for sleep arousal disorder detection based on the interaction of physiological signals and metaheuristic learning. *Computational Intelligence and Neuroscience*, 2023, 2023.

Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. Does the whole exceed its parts? the effect of ai explanations on complementary team performance. In *Proceedings of the 2021 CHI conference on human factors in computing systems*, pages 1–16, 2021.

Richard B Berry, Rohit Budhiraja, Daniel J Gottlieb, David Gozal, Conrad Iber, Vishesh K Kapur, Carole L Marcus, Reena Mehra, Sairam Parthasarathy, Stuart F Quan, et al. Rules for scoring respiratory events in sleep: update of the 2007 AASM manual for the scoring of sleep and associated events: deliberations of the sleep apnea definitions task force of the American Academy of Sleep Medicine. *Journal of clinical sleep medicine*, 8(5):597–619, 2012.

Mirella Boselli, Liborio Parrino, Arianna Smerieri, and Mario Giovanni Terzano. Effect of age on EEG arousals in normal sleep. *Sleep*, 21(4):361–367, 1998.

Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z Gajos. To trust or to think: cognitive forcing functions can reduce overreliance on ai in ai-assisted decision-making. *Proceedings of the ACM on Human-computer Interaction*, 5(CSCW1):1–21, 2021.

Daphne Chylinski, Franziska Rudzik, Dorothée Coppieters 't Wallant, Martin Grignard, Nora Vandeleene, Maxime Van Egroo, Laurie Thiesse, Stig Solbach, Pierre Maquet, Christophe Phillips, et al. Validation of an automatic arousal detection algorithm for whole-night sleep eeg recordings. *Clocks & sleep*, 2(3):258–272, 2020.

Alexander Philip Dawid and Allan M Skene. Maximum likelihood estimation of observer error-rates using the EM algorithm. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 28(1):20–28, 1979.

Rahul C Deo. Machine learning in medicine. *Circulation*, 132(20):1920–1930, 2015.

Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.

Franz Ehrlich, Tony Sehr, Moritz Brandt, Martin Schmidt, Hagen Malberg, Martin Sedlmayr, and Miriam Goldammer. State-of-the-art sleep arousal detection evaluated on a comprehensive clinical dataset. *Scientific Reports*, 14(1):16239, 2024.

Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd*, volume 96, pages 226–231, 1996.

Ahmad Fawzy, Danastri Cantya Nirmala, Denaya Khansa, and Yudhistira Tri Wardhana. Ethics and regulation for artificial intelligence in healthcare: Empowering clinicians to ensure equitable and high-quality care. 2023.

Robert Fonod. DeepSleep 2.0: automated sleep arousal segmentation via deep learning. *AI*, 3(1): 164–179, 2022.

Karl A Franklin and Eva Lindberg. Obstructive sleep apnea is a common disorder in the population—a review on the epidemiology of sleep apnea. *Journal of thoracic disease*, 7(8):1311, 2015.

Susanne Gaube, Harini Suresh, Martina Raue, Alexander Merritt, Seth J Berkowitz, Eva Lermer, Joseph F Coughlin, John V Guttag, Errol Colak, and Marzyeh Ghassemi. Do as AI say: Susceptibility in deployment of clinical decision-aids. npj Digital Medicine, 4 (31), 2021.

Susanne Gaube, Harini Suresh, Martina Raue, Eva Lermer, Timo K Koch, Matthias FC Hudecek, Alun D Ackery, Samir C Grover, Joseph F Coughlin, Dieter Frey, et al. Non-task expert physicians benefit from correct explainable ai advice when reviewing x-rays. *Scientific reports*, 13(1):1383, 2023.

Mohammad M Ghassemi, Benjamin E Moody, Li-Wei H Lehman, Christopher Song, Qiao Li, Haoqi Sun, Roger G Mark, M Brandon Westover, and Gari D Clifford. You snooze, you win: the physionet/computing in cardiology challenge 2018. In *2018 Computing in Cardiology Conference (CinC)*, volume 45, pages 1–4. IEEE, 2018.

Leilani H Gilpin, David Bau, Ben Z Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. Explaining explanations: An overview of interpretability of machine learning. In *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*, pages 80–89. IEEE, 2018.

Ary L Goldberger, Luis AN Amaral, Leon Glass, Jeffrey M Hausdorff, Plamen Ch Ivanov, Roger G Mark, Joseph E Mietus, George B Moody, Chung-Kang Peng, and H Eugene Stanley. PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. *circulation*, 101(23):e215–e220, 2000.

David Gunning, Eric Vorm, Yunyan Wang, and Matt Turek. Darpa's explainable ai (xai) program: A retrospective. *Authorea Preprints*, 2021.

Jyoti Gupta and KR Seeja. A comparative study and systematic analysis of XAI models and their applications in healthcare. *Archives of Computational Methods in Engineering*, 31(7):3977–4002, 2024.

Marc Hassenzahl, Franz Koller, and Michael Burmester. Der user experience (ux) auf der spur: Zum einsatz von www. attrakdiff. de. 2008.

Stefan Kraft, Andreas Theissler, Vera Wienhausen-Wilke, Philipp Walter, and Gjergji Kasneci. Comprehensive Polysomnography (CPS) Dataset: A Resource for Sleep-Related Arousal Research (version 1.0.0). PhysioNet, 2024. URL https://doi.org/10.13026/sxs0-h317.

Stefan Kraft, Andreas Theissler, Dr. Vera Wienhausen-Wilke, Philipp Walter, Gjergji Kasneci, and Hendrik Lensch. Alpec: A comprehensive evaluation framework and dataset for machine learning-based arousal detection in clinical practice. In *Proceedings of the sixth Conference on Health, Inference, and Learning*, volume 287 of *Proceedings of Machine Learning Research*, pages 395–429. PMLR, 25–27 Jun 2025. URL https://proceedings.mlr.press/v287/kraft25a.html.

Chih-Fan Kuo, Cheng-Yu Tsai, Wun-Hao Cheng, Wen-Hua Hs, Arnab Majumdar, Marc Stettler, Kang-Yun Lee, Yi-Chun Kuan, Po-Hao Feng, Chien-Hua Tseng, et al. Machine learning approaches for predicting sleep arousal response based on heart rate variability, oxygen saturation, and body profiles. *Digital Health*, 9:20552076231205744, 2023.

Hongyang Li and Yuanfang Guan. DeepSleep convolutional neural network allows accurate and fast detection of sleep arousal. *Communications biology*, 4(1):18, 2021.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

Melkamu Mersha, Khang Lam, Joseph Wood, Ali AlShami, and Jugal Kalita. Explainable artificial intelligence: A survey of needs, techniques, applications, and future direction. *Neurocomputing*, page 128111, 2024.

Sina Mohseni, Niloofar Zarei, and Eric D Ragan. A multidisciplinary survey and framework for design and evaluation of explainable ai systems. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 11(3-4):1–45, 2021.

Meike Nauta, Jan Trienes, Shreyasi Pathak, Elisa Nguyen, Michelle Peters, Yasmin Schmitt, Jörg Schlötterer, Maurice Van Keulen, and Christin Seifert. From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable ai. *ACM Computing Surveys*, 55(13s):1–42, 2023.

Cecilia Panigutti, Andrea Beretta, Fosca Giannotti, and Dino Pedreschi. Understanding the impact of explanations on advice-taking: a user study for ai-based clinical decision support systems. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pages 1–9, 2022.

Henna Pitkänen, Sami Nikkonen, Marika Rissanen, Anna Sigridur Islind, Heidur Gretarsdottir, Erna Sif Arnardottir, Timo Leppänen, and Henri Korkalainen. Multi-centre arousal scoring agreement in the Sleep Revolution. *Journal of Sleep Research*, 33(4):e14127, 2024.

Forough Poursabzi-Sangdeh, Daniel G Goldstein, Jake M Hofman, Jennifer Wortman Wortman Vaughan, and Hanna Wallach. Manipulating and measuring model interpretability. In *Proceedings of the 2021 CHI conference on human factors in computing systems*, pages 1–52, 2021.

Drew Prinster, Amama Mahmood, Suchi Saria, Jean Jeudy, Cheng Ting Lin, Paul H Yi, and Chien-Ming Huang. Care to explain? ai explanation types differentially impact chest radiograph diagnostic performance and physician trust in ai. *Radiology*, 313(2):e233261, 2024.

Shahbaz Rezaei and Xin Liu. Explanation space: A new perspective into time series interpretability. *arXiv preprint arXiv:2409.01354*, 2024.

Yao Rong, Tobias Leemann, Thai-trang Nguyen, Lisa Fiedler, Peizhu Qian, Vaibhav Unhelkar, Tina Seidel, Gjergji Kasneci, and Enkelejda Kasneci. Towards human-centered explainable ai: user studies for model explanations. 2022.

Rikard Rosenbacke, Åsa Melhus, Martin McKee, and David Stuckler. How explainable artificial intelligence can increase or decrease clinicians' trust in ai applications in health care: systematic review. *Jmir Ai*, 3:e53207, 2024.

Ville Satopaa, Jeannie Albrecht, David Irwin, and Barath Raghavan. Finding a" kneedle" in a haystack: Detecting knee points in system behavior. In *2011 31st international conference on distributed computing systems workshops*, pages 166–171. IEEE, 2011.

Philipp Schmidt and Felix Biessmann. Calibrating human-ai collaboration: Impact of risk, ambiguity and transparency on algorithmic bias. In *Machine Learning and Knowledge Extraction: 4th IFIP TC 5, TC 12, WG 8.4, WG 8.9, WG 12.9 International Cross-Domain Conference, CD-MAKE 2020, Dublin, Ireland, August 25–28, 2020, Proceedings 4*, pages 431–449. Springer, 2020.

Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *International conference on machine learning*, pages 3145–3153. PMlR, 2017.

Ilija Šimić, Eduardo Veas, and Vedran Sabol. A comprehensive analysis of perturbation methods in explainable AI feature attribution validation for neural time series classifiers. *Scientific Reports*, 15(1):26607, 2025.

Robert R Sokal and Charles D Michener. A statistical method for evaluating systematic relationships. 1958.

Eric J Topol. High-performance medicine: the convergence of human and artificial intelligence. *Nature medicine*, 25(1):44–56, 2019.

Joe H Ward Jr. Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, 58(301):236–244, 1963.

Thomas-Christian Wetter, Roland Popp, Michael Arzt, and Thomas Pollmächer. *ELSEVIER ESSENTIALS Schlafmedizin: Das Wichtigste für Ärzte aller Fachrichtungen*. Elsevier Health Sciences, 2012.

Yunguo Yu, Cesar A Gomez-Cabello, Syed Ali Haider, Ariana Genovese, Srinivasagam Prabha, Maissa Trabilsy, Bernardo G Collaco, Nadia G Wood, Sanjay Bagaria, Cui Tao, et al. Enhancing clinician trust in ai diagnostics: A dynamic framework for confidence calibration and transparency. *Diagnostics*, 15(17):2204, 2025.

Hasan Zan and Abdulnasır Yildiz. Multi-task learning for arousal and sleep stage detection using fully convolutional networks. *Journal of Neural Engineering*, 20(5):056034, 2023.

A Table of Notation

Table 17 lists the notation used in the main text.

Table 17: Table of notation

Symbol	Meaning
$t_{k,i}$	Start time of event i annotated by annotator k
C_j	Cluster j of temporally close annotations
ε	Temporal distance threshold for clustering
S	Kneedle sensitivity parameter for knee detection
m^{Cluster}	Minimum annotations required to form a cluster
$A_{k,j}$	Indicator that annotator k contributed to cluster j
s_k	Sensitivity (true positive rate) of annotator k
p_k	Specificity (true negative rate) of annotator k
P_j	Probability that cluster j contains a true event
P_j^{prev}	Prior event probability for cluster j from previous iteration
$\log(L_j)$	Log-likelihood under event-present hypothesis for cluster j
$\log(M_j)$	Log-likelihood under event-absent hypothesis for cluster j
D_{1j}	Unnormalized log-posterior for event present in cluster j
D_{2j}	Unnormalized log-posterior for event absent in cluster j
τ	Threshold for selecting consensus events ($P_j \geq \tau$)
$F1$	F1 score (harmonic mean of precision and recall)
$F2$	F2 score (recall-weighted F-score)
$TP,\ FP,\ FN$	True positives, false positives, false negatives
HU	Human solo
AI	AI solo
HU + AI	Human-AI collaboration
$\mathcal{B}$	Benefit ratio: $\frac{F1^{\text{HU+AI}} - F1^{\text{HU}}}{F1^{\text{AI}} - F1^{\text{HU}}}$
$\mathcal{R}$	Relative F1 score: $\frac{F1_{\text{HU+AI}}}{F1_{\text{AI}}}$
C_{GT}	Ground-truth arousal count per recording
C_x	Arousal count of source $x \in \{\text{AI, HU+AI, HU}\}$
$D_{x \to \text{GT}}$	Absolute deviation $\lvert C_x - C_{\text{GT}} \rvert$
$A_{x,\text{GT}}$	Count accuracy $\frac{1}{1+\text{APE}_{x,\text{GT}}}$
$\text{APE}_{x,\text{GT}}$	Absolute percentage error $\frac{D_{x \to \text{GT}}}{\max\{C_{\text{GT}}, \epsilon\}}$
R_{GT}	AI-baseline improvement ratio
y_{RGT}	Log improvement ratio $\log R_{\text{GT}}$
PE	Percentage error $\frac{C_{\text{HU+AI}} - C_{\text{GT}}}{\max\{C_{\text{GT}}, \epsilon\}}$

B Comparison of Feature Attribution Methods

This appendix provides an analysis of post-hoc feature attribution methods applied to a randomly selected arousal event from subject S2 (selected within the white-box quality-control regime; cf. Table 1). Our goals are to: (i) qualitatively compare local attributions across input channels and methods, (ii) quantify cross-method agreement using multiple similarity measures, and (iii) relate local agreement to global channel importance rankings derived from aggregated attributions.

Methods in Brief We employ two gradient-based explainability methods to generate attributions for our neural time-series model. DeepLIFT is our primary method for the user study (see

Section 2.2). It attributes relevance by contrasting activations to a reference activation in a computationally efficient manner. GradientSHAP, used here as a complementary method, builds on the SHAP framework by using gradients and random perturbations to estimate feature attributions (Lundberg and Lee, 2017). It approximates SHAP values by averaging gradients taken along random paths from reference inputs (baselines) to the target input. In each of many trials it (i) slightly perturbs the input, (ii) randomly selects a baseline, (iii) picks a random point on the straight line between that baseline and the (noisy) input, and (iv) computes the target output's gradient there. Averaging over trials yields SHAP values that are essentially the expected gradient weighted by the input – baseline differences. We use the standard implementation from the `captum` library and use the same baselines as for DeepLIFT (see Section 3.4.1).

Local Attributions Figure 21 displays the per-channel attributions for DeepLIFT and GradientSHAP for a 60-second window centered on the model's predicted event onset (dashed gray line).

This matches the user-study visualization with *low* threshold settings which reveal the most fine-grained local contributions (see Section 3.4.1). Importantly, these thresholds influence the computation of global channel relevance but not the local cross-method similarity analysis, which uses all available attributions.

To visually compare temporal attributions directly and prepare for the correlation analysis, Figure 22 overlays normalized DeepLIFT and GradientSHAP traces for the same channels ordered by Spearman correlation over the normalized signals. Normalization is performed per channel using min-max scaling with an exclusion window of $\pm 0.5\,\mathrm{s}$ around the prediction point to prevent spike domination at $t = 0$. This window is *not* removed from the signals used for the correlation analysis, it only affects the normalization parameters.

The ordering by ρ reflects which channels display most similar attribution patterns (RIP.Abdom, EMG). Visually, high-ρ channels show aligned bursts across methods, while lower-ρ channels exhibit timing or spread differences, indicating method-specific sensitivities. Based on Spearman ρ, the correlation for most channels is weak, while one channel (RIP.Abdom) reaches moderate correlation.

Before proceeding, we test the robustness of our approach by comparing alternative correlation measures for quantifying agreement between the two attribution methods. Figure 23 summarizes agreement between the two attribution methods across channels. We report Spearman rank correlation (monotonic agreement), Pearson correlation (linear agreement), and cosine similarity (directional alignment).

Panels A, B and E reveal that Pearson correlation and cosine similarity typically rank channels similarly, reflecting shared sensitivity to overall shape and directional alignment. Spearman correlation, by focusing on order rather than magnitude, penalizes localized discrepancies (e.g., minor shifts in peaks or differing spike widths). Because our interest is whether methods agree on *when* a channel is relevant (focusing on the temporal structure of the attributions) rather than exact amplitude scaling, we use *Spearman correlation* as the primary agreement measure. This choice is further supported by Figure 22, where channels with high ρ consistently exhibit similar patterns across methods.

Global Attributions We next relate local agreement to global feature importance. We compute global relevance by aggregating the positive attributions for each channel over time at *low* threshold selection.

Figure 24 combines two complementary views. Panel A shows the correlation of global relevance rankings between DeepLIFT and GradientSHAP across all channels. Panel B displays a slope graph for the top-15 channels (by DeepLIFT relevance), comparing their DeepLIFT relevance rank, GradientSHAP relevance rank, and agreement-score (Spearman) rank.

Panel A demonstrates a *very strong* correlation between global relevance rankings from both methods, indicating robust agreement on which channels matter most overall. Panel B shows that (i) the top-three channels are consistently ranked highly by both approaches, and (ii) a small number of channels (e.g., "Schnarc") deviate substantially between methods, or are globally important but show relatively low local attribution agreement (e.g., "Pulse"). This pattern shows that, even when global importance aligns, temporal attribution structure can still differ.

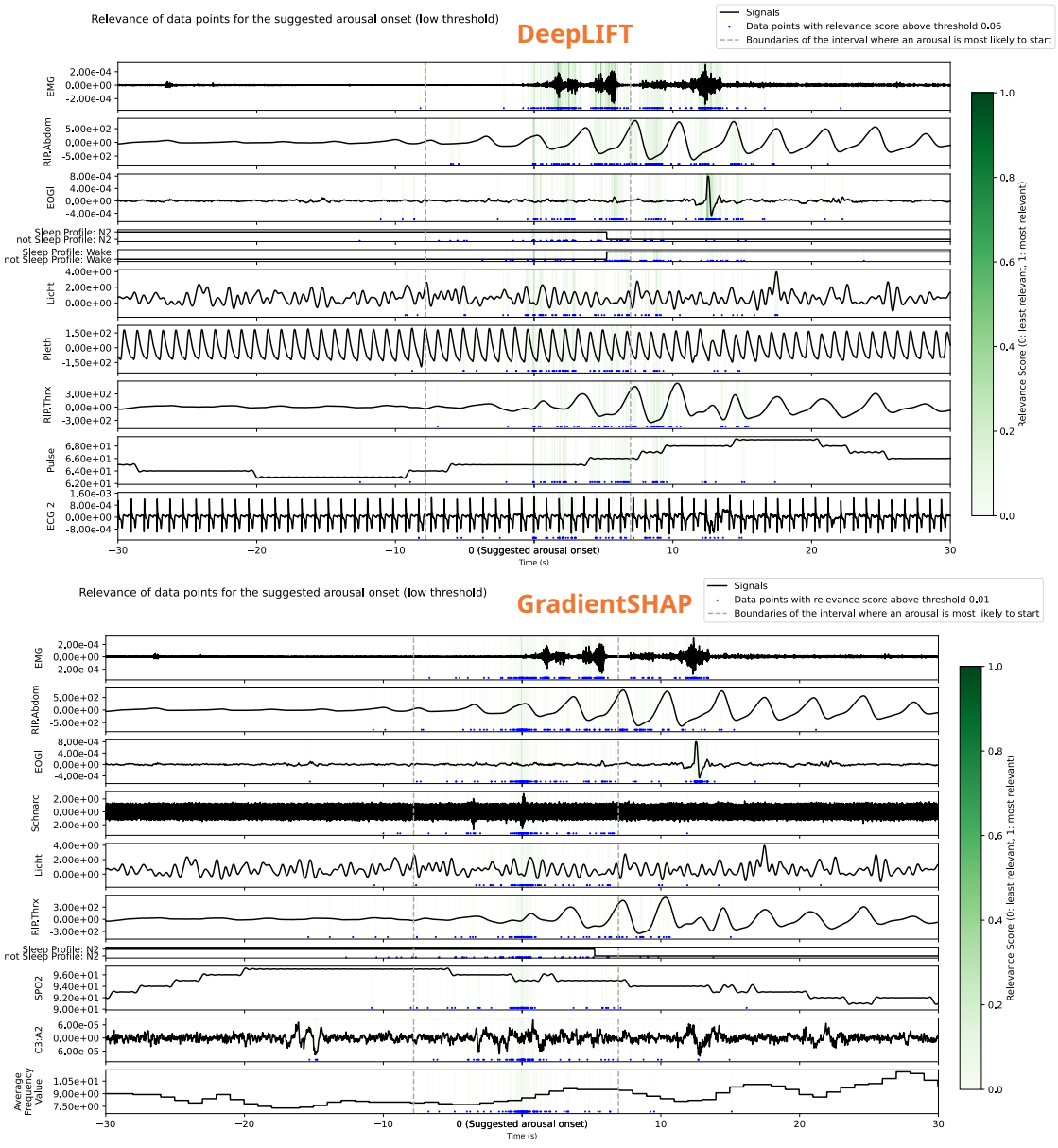


Figure 21: **Local feature attributions (DeepLIFT vs. GradientSHAP).** Both methods highlight consistent event-proximal relevance spikes in several respiratory and EMG-derived channels, while also revealing method-specific differences in sparsity and spread of relevance.

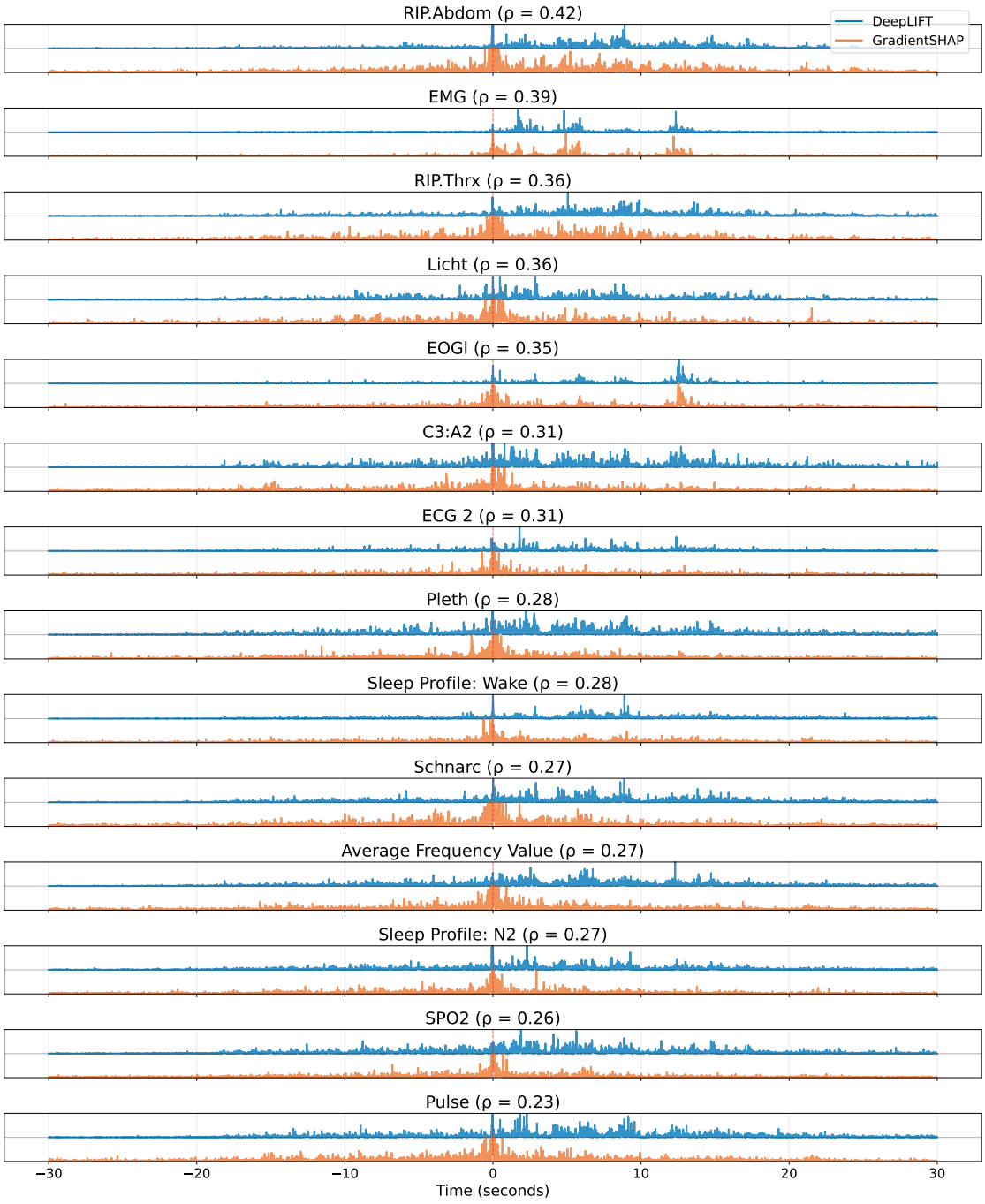


Figure 22: **Side-by-side temporal comparison of normalized attributions (DeepLIFT vs. GradientSHAP).** Channels are ordered by cross-method Spearman correlation ρ. Normalization excludes ± 0.5 s around the predicted onset to avoid peak inflation. The correlations are computed on the full signals, not the masked signals used for normalization.

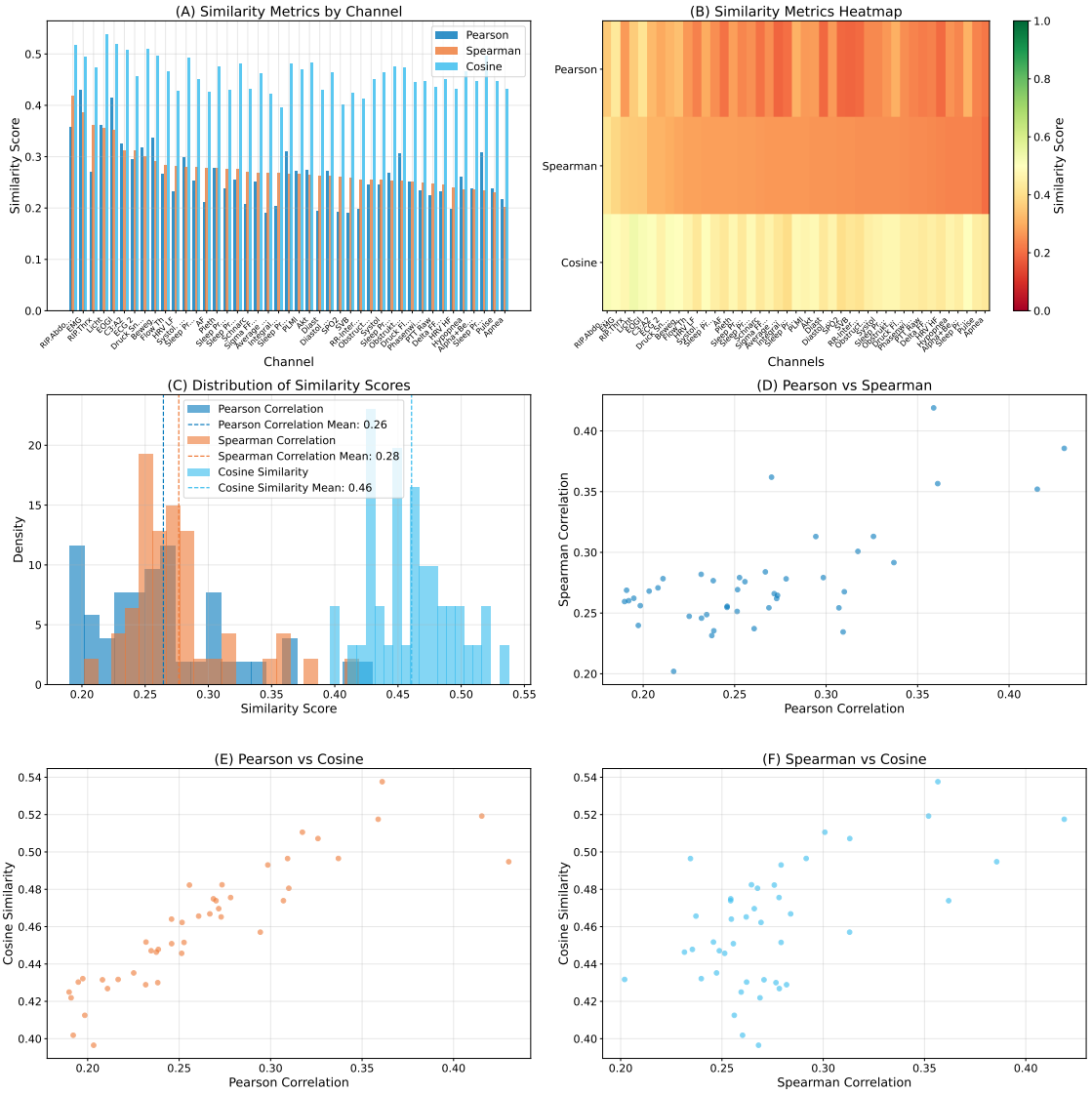


Figure 23: **Similarity between methods across channels for three different correlation measures.** Plots (A) and (B) show corresponding channel-wise similarity scores ordered by Spearman correlation. Pearson correlation and cosine similarity tend to co-vary, while Spearman correlation differs markedly, emphasizing monotone agreement in temporal structure.

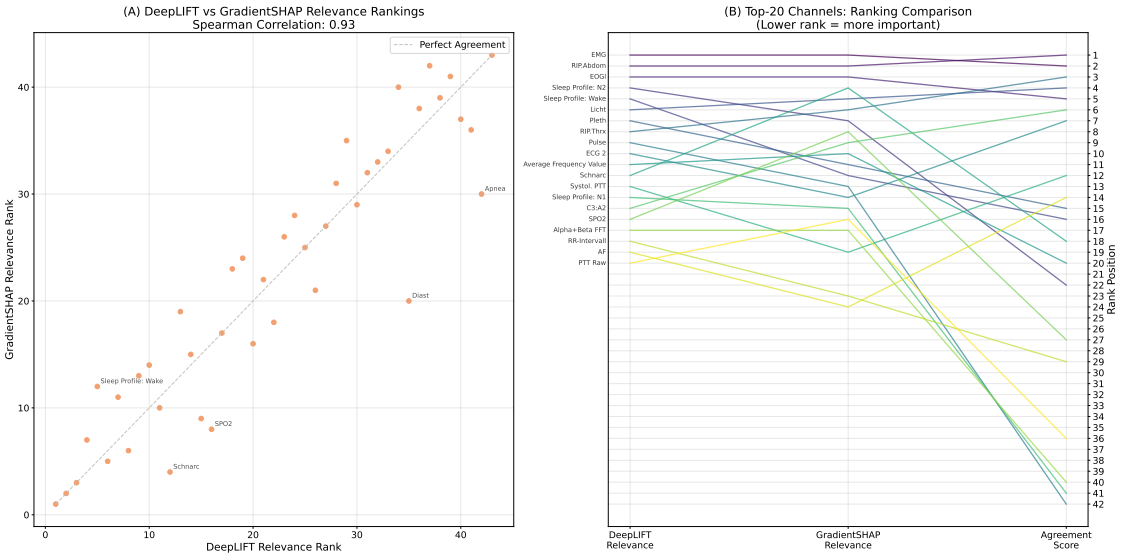


Figure 24: **Global relevance and cross-method correspondence.** Panel A shows a very strong rank correlation between DeepLIFT and GradientSHAP global relevance across channels. Panel B depicts a slope graph for the top-20 channels by DeepLIFT relevance: both methods agree on the leading channels, while a few channels show notable divergence or high global relevance but comparatively low local agreement.

Discussion Overall, DeepLIFT and GradientSHAP provide a largely coherent view of feature attribution patterns in this exploratory analysis. Notably, both methods identify the same three channels (EMG, RIP.Abdom, and EOGl; see Figures 21 and 24B) as most important globally. These channels also demonstrate the highest level of cross-method agreement, as measured by Spearman correlation. Although the agreement scores for these channels are only moderate or on the borderline of moderate, visual inspection (Figure 24B) reveals a clear and meaningful alignment in their overall temporal attribution patterns. Conversely, channels deemed less important by both methods (e.g., Pulse or Average Frequency Value) tend to show lower agreement scores and more pronounced visual discrepancies in attribution patterns (see Figure 22). In fact, the time-series for Pulse and Average Frequency Value do not display any distinctive features from a layman's perspective in contrast to e.g. EMG, RIP.Abdom, or EOGl (see Figure 21), which may help explain the lack of consistent attribution. Importantly, the global rank correlation across all channels remains very strong ($\rho = 0.93$), underscoring robust overall concordance between the two attribution methods. Of particular note is the "Pulse" channel, which demonstrates an intriguing discrepancy: while it ranks as the second most globally relevant channel according to DeepLIFT across the entire dataset (see Figure 3), it shows the lowest consistency in its attribution pattern when compared to GradientSHAP (see Figure 24B). Determining whether this inconsistency is merely an anomaly specific to the event examined here, or indicative of a broader phenomenon, would require more extensive analysis across a larger set of events. Such an investigation, as well as the evaluation of additional post-hoc attribution methods, remains a valuable direction for future work but lies beyond the scope of this study.

Collectively, these findings support the robustness of our choice of *DeepLIFT* for the user study: It delivers clear, interpretable local explanations and demonstrates good concordance with GradientSHAP, showing clear visually consistent patterns of agreement for the most important channels at the local level and excellent consistency in global relevance across channels.

C Mapping of Subject IDs

Table 18 shows the mapping of subject IDs, as used in Table 1, to file IDs as specified in the documentation of the CPS dataset (Kraft et al., 2024).

Table 18: **Mapping of Subject IDs to File IDs**. Subject IDs from Table 1 are mapped to the corresponding file IDs as specified in the CPS dataset documentation (Kraft et al., 2024).

Subject-ID	File-ID
S1	FLsgQZoIGHx1G3LmdD7jtICMik2EKRKN
S2	Su02hndUSYGSKJmcSqroKmtDjXIJ4y60
S3	kKDzUlAprXqDz84Nrw9UP1W0jpgUKkhN
S4	YFQX33c8EEoapTndd2084KbUuUmtj7xF
S5	LcsapTberZwzU7qyEr11andO59HTOVCv
S6	tIgyhF8T1BOZnu7h6jb58igU5MAGgdo9
S7	CzwqE37s81YahjNICSXI2Tb4Fmp6bclE
S8	vHJMSYFIl1TfLweQ5DWMGN5f47ULFNxe
S9	leySrSnra9yAO3eTJIGB55nrjRS3RqIW
S10	KB84bUmLWOrKCKkISCn8QuNBhF5mg0L8
S11	MT1zW5iB0h1bxF42QBpyDqotQk7NcnHw
S12	RpARZ17l5osnFUcqIj2aTOsgRBMDutoA
S13	mXHZZ887A9fcZgOmnxhnPVHwu5ECljDG
S14	oEJ7fslCTL7s0OfIe7nYIPqo7Il4rMjI

D Construction of Consensus Ground-Truth

Figure 25 shows the human annotations over multiple time ranges for the two manually annotated samples S2 and S4 (see Table 1). The annotations are clustered using Agglomerative Clustering (see Section 2.3.1). The consensus annotations are constructed using expectation maximization (see Section 2.3.2).

E Free-Text Responses

This section contains questions with free-text responses from the participants, translated into English. If questions refer to former questions, the reference is given as well.

Describe your experiences with the web app. What did you like and what could be improved?

1. Keyboard shortcuts
   Full screen utilization
2. Training beforehand would be good
   Then handling becomes easier over time
3. Good signal resolution, many adjustment options for evaluation
4. It is very comprehensive and individually adjustable but the many submenus make it very complicated. The expandable menus are annoying. The "Event created" info is unnecessary.
5. Initially it seems complicated. Once you get an overview it is very clear. Very detailed and professionally created.
6. Good
7. I find the web app user-friendly.

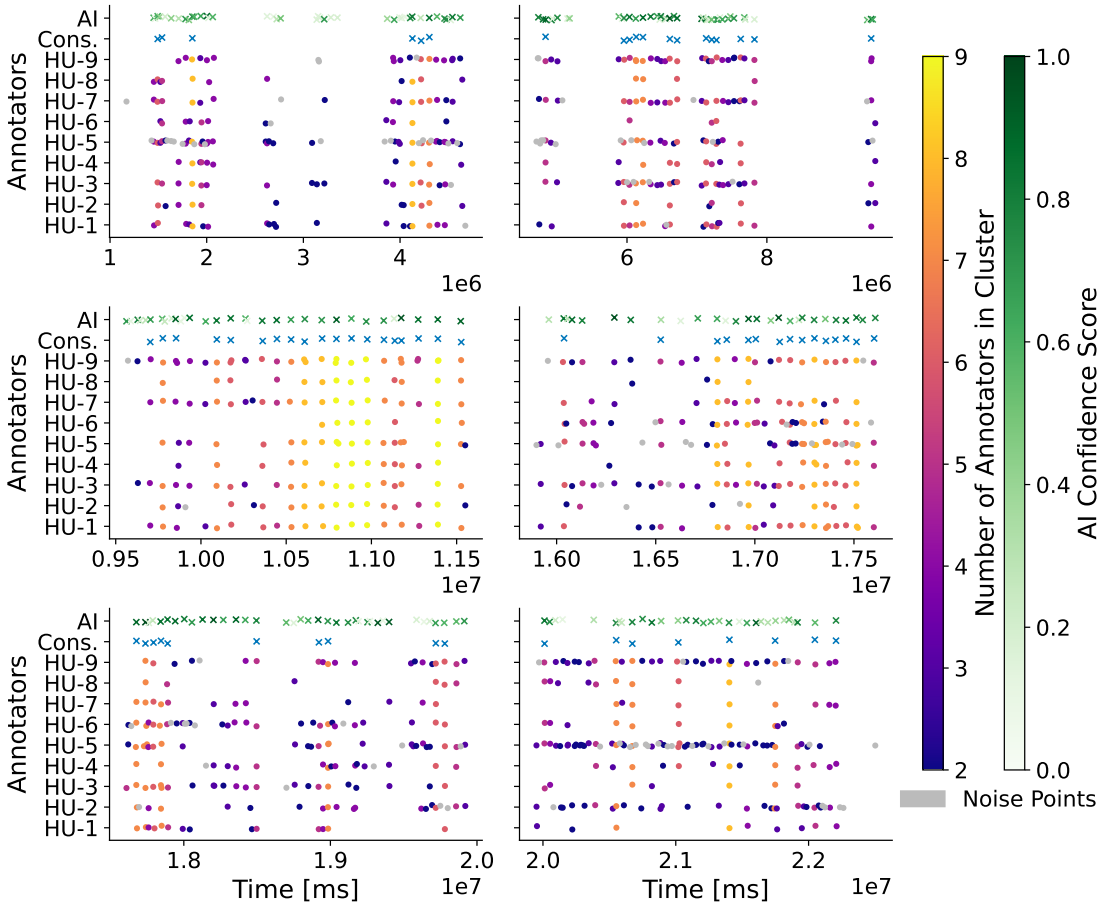


Figure 25: **Faceted Timeline Plot of Clustering Results using Agglomerative Clustering**. The plot displays human solo annotations across multiple panels, with data from patient samples *S2* and *S4* concatenated into approximately six hours of recording time. Annotations of the same color and vertical alignment belong to the same cluster. The horizontal displacement is for visual clarity only. Large intervals without annotations have been omitted. Consensus annotations (*Cons.*) are marked as blue crosses, while AI annotations are marked as green crosses, with the hue representing the AI model's confidence score. The color scale on the right indicates the number of annotators in each cluster. Annotator *HU-9* represents the human annotations used as the CPS ground truth for AI model training.

8. More additional markings desired

Here is space for explanations of your choice in the last question.

Reference: Could you imagine using the present system for arousal evaluation in the sleep laboratory?

1. Multiple relevant channels at a glance would be better

Here is space for explanations of your choice in the last question.

Reference: Which approach do you prefer in connection with AI support: (1) You have access to the AI-determined arousals and explanations from the beginning or (2) You first evaluate manually without AI support and use the AI support for quality control afterwards? Does your preference differ depending on whether the AI support is a Black Box or transparent?

1. One actually gets influenced and questions one's decisions again

Has the support from the Black Box AI influenced your decisions? If yes, please give examples. [Other]

1. Regarding confidence
2. Checked own decision again when no arousal was given
3. one or the other arousal
4. Generally
5. positive

Do you have suggestions for improving the Black Box AI support? [Other]

1. a bit faster
2. Programming of AASM rules e.g. no arousals while awake
3. limitation of the marked area to the blue line

Here is space for explanations of your choice in the last question.

Reference: How did your approach differ between the three modes: (1) manual evaluation of arousal events, (2) evaluation with Black Box AI support and (3) evaluation with transparent White Box AI support?

1. Decisions become plausible through the other channels and simplify the selection

Justify your answer to the last question.

Reference: Did the explanations for the AI-determined arousals help you with their validation?

1. The confidence and probability give me some more insight
2. In case of uncertainty it definitely helped and was often conclusive
3. it supports and you take a closer look at it again
4. Due to lack of space, additional channels could simply be viewed in the explanations, especially when the AI has already classified them as relevant, they are very interesting.
5. The explanations helped but were not decisive in my decisions
6. Explanation was the confirmation for me
7. For arousals where I wasn't quite sure
8. Confidence

Did you gain interesting or unexpected insights from the transparent AI explanations about the AI's approach to arousal determination? If yes: What are they? [Other]

1. that the AI really uses all available data

2. The data that the AI uses for assessment differs greatly from the data that an evaluator normally pays attention to
3. Arousals are not differentiated here
4. I was surprised that AI marked arousals that I saw as awake

Could you identify one or more recurring patterns by which the AI recognizes arousals? If yes: What are they? [Other]

1. Pulse rate
2. PLMs (don't always lead to an arousal) and desaturations with breathing events
3. Heart rate increase
4. Thorax and abdomen breathing, EMG
5. EEG
6. Breathing pattern /PLM

Has the support from the transparent AI influenced your decisions? If yes, please give examples. [Other]

1. The explanation based on partially convinced me to score an arousal
2. occasionally
3. PLM seen and I didn't but she was right.
4. Generally
5. positively influenced
6. Breathing pattern /PLM

Did you find systematic errors that the AI model makes? If yes: Which ones? [Other]

1. Distance of arousals, too long events
2. Arousals sometimes too close together without enough sleep in between
3. Arousal drawn but patient awake
4. see above

Do you have suggestions for improving the transparent AI support?

1. pay more attention to EEG frequency increases and only then the other parameters.
2. better differentiation of arousal

Justify your answer to the last question.

Reference: The AI model was optimized to miss as few arousals as possible. For this reason, there are more frequently false positive determined arousals, i.e., the model marks arousals more frequently where there actually are none. For this purpose, the threshold for detecting an arousal was set relatively low. How do you find this approach?

1. Too many false arousals. I get somewhat delayed by this
2. you reflect on your evaluation again and possibly correct yourself again
3. I score an arousal rather quickly, rather than having to sight-decide and then possibly decide
4. I'd rather draw a few "forgotten" arousals myself than having to delete many again.
5. In my opinion, the AI was almost too generous with marking arousals
6. 95% correctly marked
7. Skipped
8. without justification

When the AI had correctly recognized an arousal, how good was in your opinion generally the agreement of the AI-determined range in which the arousal start likely lies with the actual start of the arousal? Was the range too large or the most likely point (vertical dashed line) too far from the actual start point? Please consider in your answer what degree of agreement is required for the practical usability of the AI support for arousal evaluation. [Other]

1. last recording the areas fit perfectly, with the others the point was a bit too far away

If the accuracy was not good enough for you in the last question: How much tighter would the predicted range have to be or how much closer would the most likely point have to be to the actual start point of an arousal?

Reference: When the AI had correctly recognized an arousal, how good was in your opinion generally the agreement of the AI-determined range in which the arousal start likely lies with the actual start of the arousal? Was the range too large or the most likely point (vertical dashed line) too far from the actual start point? Please consider in your answer what degree of agreement is required for the practical usability of the AI support for arousal evaluation.

1. The start point should be closer to the highest point of the curve, arousals all start too early
2. very different, but the start point of the arousal is very important
3. +/- 3 sec
4. the line was exactly right
5. 0.3 -0.5 factor

How did your approach differ between the three modes: (1) manual evaluation of arousal events, (2) evaluation with Black Box AI support and (3) evaluation with transparent White Box AI support?

1. 1. I thought about what the AI would rate now
   2. I missed the threshold/confidence
   3. I could best verify the AI
2. all 3 the same,
   checked the same for the given arousals
   the others afterwards
3. Setting the modes and then starting
4. Not at all, except that I thought about why the blackbox might have marked this now and with the whitebox I looked at why it marked it and then as always decided who was right.
5. With the White Box AI, the evaluation went much faster because you can rely very much on the AI having marked everything important, unlike manual evaluation
6. White Box was better for me
7. Not at all
8. transparent White Box AI support was my favorite

Which approach do you prefer in connection with AI support: (1) You have access to the AI-determined arousals and explanations from the beginning or (2) You first evaluate manually without AI support and use the AI support for quality control afterwards? Does your preference differ depending on whether the AI support is a Black Box or transparent?

1. 1. is certainly easier - for experienced evaluators! (somewhat tempting for newcomers)

2. would prefer the transparent AI from the beginning
3. 1. with AI support
4. 1 and transparent
5. I prefer approach 1. Yes the preference differs
6. 2
7. From the beginning with AI
8. With AI support more confidence and speed

If you wished for improvements in the last question or would not use the system, justify your answer.

Reference: Could you imagine using the present system for arousal evaluation in the sleep laboratory?

1. as already mentioned above
2. shorter markings
3. speed
4. The user interface would need to be simplified. Ideally, the AI would integrate into existing evaluation systems.
5. An option to set preferences regarding how strict the AI is when marking arousals would be very helpful.
6. more color markings

Would you participate in a similar study again in the future? Why or why not?

1. Yes
2. yes, was very interesting
3. yes, but crucial is who guides you through the study
4. Yes, of course.
5. Yes. Professional development.
6. Yes
7. I would like to participate in a similar study
8. Yes