Self-Explainable Machine Learning for Medical Image Analysis

Dissertation

der Mathematisch-Naturwissenschaftlichen Fakultät

der Eberhard Karls Universität Tübingen

zur Erlangung des Grades eines

Doktors der Naturwissenschaften

(Dr. rer. nat.)

vorgelegt von

Kerol Djoumessi

aus Dschang/ Kamerun

Tübingen

2025

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

Tag der mündlichen Qualifikation:	30.04.2026
Dekan:	Prof. Dr. Thilo Stehle
1. Berichterstatter:	Prof. Dr. Philpp Berens
2. Berichterstatter:	Prof. Dr. Jakob Macke

Abstract

Artificial intelligence (AI), particularly deep neural networks (DNNs), has driven major advances across various sectors, including education, transportation, finance, entertainment, and communication. Yet in high-stakes domains such as healthcare, adoption remains limited due to their black-box nature, which obscures decision-making processes and hinders human interpretation. This lack of transparency undermines trust, restricts clinical use, and prevents clinicians from validating predictions. To mitigate this issue, post-hoc attribution methods attempt to generate visual explanations that approximate model reasoning. However, these approaches are often unreliable in medical imaging, failing to reflect the model's true decision process faithfully, and are vulnerable to spurious correlations. While inherently interpretable or self-explainable models embed explanations directly into their architecture, they often trade off accuracy, offer limited transparency, and lack generalizability or quantitative evaluation. Thus, transforming black-box models into self-explainable systems without sacrificing classification predictive performance remains a key challenge.

This thesis addresses these challenges through three main contributions. First, we introduce Sparse BagNet, a self-explainable DNN built upon BagNet—that already provides patchwise local explanations—and further enhances transparency by removing the average pooling layer and replacing the classification layer with a convolutional layer. This modification produces class evidence maps that preserve spatial information, while a lasso penalty enforces sparse explanations. Evaluated in a retrospective clinical study, Sparse BagNet's explanations improved ophthalmologist' diagnostic accuracy by 17% while reducing their decision time by approximately 25%.

Second, to extend local explanation toward global interpretability, we developed ProtoBagNet, which combines BagNet's small receptive fields with prototype learning. ProtoBagNet provides both local explanations through prototype similarity maps and global explanations via learned prototypes. By incorporating a dissimilarity loss, it encourages diverse and non-redundant prototypes, overcoming limitations of prior prototype-based models and producing more precise, faithful explanations that better capture the model's underlying reasoning.

Finally, we generalized the Sparse BagNet into SoftCAM, a protocol for converting standard convolutional neural networks (CNNs) into self-explainable models. Like Sparse BagNet, SoftCAM systematically replaces the average pooling and fully connected layers with a convolutional classifier, but extends the sparsity regularization from Lasso to ElasticNet, allowing explanations to adapt to dataset-specific characteristics. Evaluated on several medical imaging datasets against established post-hoc attribution methods, SoftCAM consistently produced more precise and faithful explanations while maintaining performance comparable to black-box baselines. Building on this framework, we further designed a fully convolutional hybrid CNN-Transformer architecture for retinal disease detection, combining the locality of convolution with the long-range dependency modeling of transformers while preserving inherent interpretability.

Together, these contributions advance the development of transparent, trustworthy, and clinically useful AI systems, while establishing rigorous standards for evaluating model explainability, with principles that can be extended beyond medical imaging to other high-stakes vision tasks.

Zusammenfassung

Künstliche Intelligenz (KI), insbesondere künstliche neuronale Netzwerke, hat zu großen Fortschritten in verschiedenen Sektoren geführt, darunter Bildung, Transport, Finanzen, Unterhaltung und Kommunikation. In anderen Bereichen, wie zum Beispiel dem Gesundheitswesen, werden diese Technologien jedoch noch immer zurückhaltend eingesetzt, da es oft schwer ist, die Entscheidungsprozesse von Diagnosen nachzuvollziehen. Diese fehlende Transparenz untergräbt das Vertrauen in KI-Systeme, limitiert den klinischen Einsatz und macht es Ärzten schwer, die Vorhersagen der Modelle zu überprüfen. Um dieses Problem zu lösen, wurden Attributionsmethoden entwickelt, die post-hoc visuelle Erklärungen generieren, um den Entscheidungsprozess nachvollziehbar zu machen. In der medizinischen Bildgebung erweisen sich diese Ansätze jedoch häufig als unzuverlässig. Sie spiegeln nicht wirklich wider, wie das Modell tatsächlich zu seiner Entscheidung gelangt, und neigen dazu, simple Korrelationen als Begründungen hervorzuheben. Zwar gibt es auch Modelle, die von Grund auf dafür konzipiert sind, interpretierbar zu sein und Erklärungen direkt über ihre Architektur zu integrieren, doch diese gehen oft zulasten der Genauigkeit, bleiben in ihrer Transparenz beschränkt und lassen sich weder gut auf andere Problemstellungen übertragen noch quantitativ bewerten. Die zentrale Herausforderung besteht daher darin, aus schwer zu interpretierenden Modellen selbsterklärende Systeme zu machen, ohne dabei Einbußen bei der Genauigkeit von Klassifikationen hinnehmen zu müssen.

Die vorliegende Dissertation begegnet diesen Herausforderungen mit drei wesentlichen Ansätzen. Zunächst wird das Sparse BagNet vorgestellt, ein selbsterklärendes neuronales Netz, das auf der BagNet-Architektur aufbaut. Das Sparse BagNet verbessert die bereits vorhandenen lokalen Erklärungen von BagNet, indem das sogenannte Pooling im Netzwerk entfernt wurde und die finale Klassifikation im Netzwerk durch eine Faltungsschicht ersetzt wird. Dadurch entstehen visuelle evidenzbasierte Abbildungen der medizinischen Bilder, während eine Lasso-Regularisierung dafür sorgt, dass die Felder für die Erklärung lokal und sparse sind. In einer retrospektiven klinischen Studie konnte gezeigt werden, dass die Erklärungen von dem Sparse BagNet die diagnostische Genauigkeit von Augenärzten um 17% steigerten und gleichzeitig deren Entscheidungszeit um etwa 25% verkürzten.

Die zweite Studie erweitert die lokale Erklärbarkeit um eine globale Perspektive: Das ProtoBagNet kombiniert die kleinen rezeptiven Felder von BagNet mit prototypbasiertem Lernen. Dadurch liefert ProtoBagNet sowohl lokale Erklärungen mittels Abbildungen, die Ähnlichkeiten zu Prototypen anzeigen, als auch globale Erklärungen durch die gelernten Prototypen selbst. Eine spezielle Verlustfunktion, die sich auf Unähnlichkeiten zwischen Bildern konzentriert, sorgt dafür, dass die Prototypen vielfältig und nicht redundant sind (ein Problem, das frühere prototypbasierte Ansätze plagten) und führt so zu präziseren Erklärungen.

Die dritte Studie verallgemeinert den Ansatz des Sparse BagNet zur SoftCAM-Anwendung. Mit dieser lassen sich herkömmliche konvolutionale künstliche neuronale Netzwerke in selbsterklärende Modelle umwandeln. Ähnlich wie das Sparse BagNet ersetzt auch SoftCAM systematisch den Pooling-Schritt im Netzwerk durch einen konvolutionalen Klassifikator. Zusätzlich wird in SoftCAM jedoch auch die Lasso-Regularisierung durch ein ElasticNet ersetzt, welches es erlaubt, die Erklärungen besser an die spezifischen Eigenschaften verschiedener Datensätze anzupassen. In Evaluationen auf mehreren medizinischen Bildgebungsdatensätzen lieferte SoftCAM durchweg präzisere und verlässlichere Erklärungen als etablierte Post-hoc-Methoden, während es eine

vergleichbare Genauigkeit wie klassische nicht-selbsterklärende Modelle liefert. Darauf aufbauend wurde eine Methode entwickelt, die sich vollständig auf konvolutionale Netzwerke und Transformer stützt. Diese neue Methode kombiniert die Vorzüge von konvolutionalen künstlichen Netzwerken mit Transformern und kann globale Abhängigkeiten modellieren, ohne dabei die inhärente Interpretierbarkeit zu verlieren.

Zusammengefasst tragen die genannten Studien zur Entwicklung transparenter, vertrauenswürdiger und klinisch nutzbarer KI-Systeme bei und etablieren strenge Bewertungsstandards für die Modellerklärbarkeit. Dabei ist die Anwendbarkeit auch über die medizinische Bildgebung hinaus zu denken und lässt sich auf andere sicherheitskritische Anwendungen, wie zum Beispiel im Bereich Computer Vision, übertragen.

Acknowledgements

My PhD has been an exciting and transformative journey, during which I had the privilege of being part of a diverse and welcoming research community. This experience allowed me to collaborate with and learn from many inspiring people, while building meaningful friendships along the way.

I would first like to express my deepest gratitude to my parents, Djoumessi Voukeng Celestin and Nguefack Jenevieve, for their unconditional love, endless support, and unwavering belief in me. The values you instilled and the strength you provided have shaped who I am today. Your constant encouragement and the safe foundation you created gave me the confidence to explore the world fearlessly and pursue my aspirations. I am profoundly thankful for your love, understanding, and enduring presence throughout this journey.

I am deeply grateful to my supervisor, Prof. Dr. Philipp Berens, for his invaluable guidance, support, and encouragement throughout my doctoral research, while granting me the freedom to pursue the research questions that inspire me. His expertise, insightful feedback, and unwavering dedication have been instrumental in shaping this thesis. I am also deeply grateful for his kindness and for fostering a welcoming and supportive environment for an international research team.

My heartfelt thanks also go to my co-supervisor, Prof. Dr. Jakob Macke, as well as Prof. Dr. Lisa Koch and Dr. Bubacarr Bah, for their guidance, mentorship, and valuable inputs throughout this research endeavor. I would also like to thank thank Dr. Wieland Brendel for serving on my TAC committee and for providing constructive feedback that has greatly contributed to the development of this work.

I had the great pleasure of working with Indu Ilanchezian, Samuel Mensah, Julius Gervelmeyer, Sarah Müller, Ziwei Huang, and all members of the Hertie Institute for AI in Brain Health. I am deeply grateful for their camaraderie, support, and active engagement in numerous fruitful discussions and collaborations, which have enriched my research experience. Special thanks go to the ophthalmologists Dr. Hanna Faber, Dr. Annekatrin Rickmann, Dr. Natalia Simon, Dr. Laura Kühlewein for their time, invaluable contributions, and insightful feedback, which have enhanced the significance of this research.

Many thanks to Sebastian Damrich, Jan Lause, Ziwei Huang, Sacha Sokoloski, and Lisa Schmors for reviewing this work and providing constructive feedback. I am also thankful to Valeska Botzenhardt and Chiu Yi Lam for their help and support with administrative tasks.

I would like to express my sincere appreciation for the invaluable support and funding provided by the International Max Planck Research School for Intelligent Systems (IMPRS-IS). Special thanks go to Leila Masri and Sara Sorce for their motivational discussions and for creating a lively and supportive atmosphere within the IMPRS-IS doctoral program.

I am also grateful to my siblings Nadege, Baker, Pinos, Gina, Lucrene, Line, Ronsard, and Hans, as well as my friends Abdul, Yvan, Jerry, Xavier, Eric, Gilles, and Macqueen for their unwavering support, encouragement, and motivation throughout this journey. Your presence and belief in me have been a constant source of strength and inspiration.

Finally, to my son, Djoumessi Voukeng Christ Keryan, you have been my greatest source of motivation, and I am forever grateful to have you in my life.

Thank you!

Contents

List of Figures

List of Acronyms

ACE Automated Concept-based Explanations

AI Artificial Intelligence

AIA Artificial Intelligence Act

AMD Age-Related Macular Degeneration

AUC Area Under the Curve

CAM Class Activation Map

CNN Convolutional Neural Network

CvT Convolutional vision Transformers

DL Deep Learning

DME Diabetic Macular Edema

DNN Deep Neural Network

DR Diabetic Retinopathy

ERM Epiretinal Membrane

FCL Fully Connected Layer

GAP Global Average Pooling

GDPR General Data Protection Regulation

LIME Local Interpretable Model-agnostic Explanations

MHSA Multi-Head Self-Attention

ML Machine Learning

OCT Optical Coherence Tomography

SA Self-Attention

SGD Stochastic Gradient Descent

SHAP Shapley Additive Explanations

TCAV Testing with Concept Activation Vectors

ViT Vision Transformer

XAI Explainable Artificial Intelligence

List of publications

Publications

Kerol Djoumessi, Indu Ilanchezian, Laura Kühlewein, Hanna Faber, Christian F Baumgartner, Bubacarr Bah, Philipp Berens, and Lisa Koch. "Sparse activations for interpretable disease grading". In *Medical Imaging with Deep Learning*, 2023.

Kerol Djoumessi, Ziwei Huang, Laura Kühlewein, Annekatrin Rickmann, Natalia Simon, Lisa Koch, and Philipp Berens. "An inherently interpretable AI model improves screening speed and accuracy for early diabetic retinopathy". *PLOS Digital Health*, 2025.

Kerol Djoumessi, Bubacarr Bah, Laura Kühlewein, Philipp Berens, and Lisa Koch. "This actually looks like that: Proto-bagnets for local and global interpretability-by-design". In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 718–728, 2024.

Kerol Djoumessi, Samuel Mensah, and Philipp Berens. "A Hybrid Fully Convolutional CNN-Transformer Model for Inherently Interpretable Disease Detection from Retinal Fundus Images". *Accepted and presented at the International Workshop on Interpretability of Machine Intelligence in Medical Image Computing (IMIMIC) at MICCAI*, 2025.

Preprint

Kerol Djoumessi and Philipp Berens. "Soft-CAM: Making black box models self-explainable for high-stakes decisions". *arXiv preprint arXiv:2505.17748*, 2025.

1 Introduction

This cumulative thesis summarizes the key contributions of my Ph.D. research, presented through five first-author papers. Three of these have been published in international peer-reviewed conferences [1, 2] and a journal [3], a fourth has been accepted for publication in an international peer-reviewed workshop [4], and the fifth is currently available as a preprint [5]. All were completed during my doctoral studies at the *Data Science Department* of the Hertie Institute for AI in Brain Health, University of Tübingen.

1.1 Overview of machine learning techniques

The rapid advancement of artificial intelligence (AI) over the past two decades has enabled machines to perform complex tasks with remarkable accuracy and efficiency [6–8]. In computer science, AI broadly refers to methods that mimic aspects of human intelligence, ranging from rule-based systems and decision trees to modern machine learning (ML) approaches [9, 10]. Machine learning, which allows computers to learn from data and generalize to unseen cases, has become one of the most transformative areas of AI [7, 11]. Within ML, deep learning (DL) based on deep neural networks (DNN) uses large data sets and computational resources to achieve state-of-the-art performance in domains such as healthcare, finance, computer vision, and natural language processing [7, 12, 13].

Machine learning methods are commonly categorized into supervised, unsupervised, and reinforcement learning [10]. Supervised learning—the primary focus of this work—trains models on labeled input-output pairs to learn mappings from data to predictions. It has been widely applied across domains: in computer vision for medical image diagnosis (e.g., detecting diabetic retinopathy from retinal fundus images [14, 15] or pneumonia from chest x-rays [16, 17]); in finance for fraud detection and customer risk assessment [7, 18], in language technology for natural language processing and speech recognition [8, 13]; and in transportation for traffic flow and congestion prediction [8, 13]. Supervised learning encompasses both classification (predicting discrete labels) and regression (predicting continuous values) [10, 13]. By contrast, unsupervised methods seek to uncover hidden patterns in unlabeled data [10, 19], while reinforcement learning focuses on agents that learn optimal strategies through interaction with an environment [10, 20].

Among ML approaches, supervised learning remains the most widely adopted [12, 18], driven by the increasing availability of labeled data across diverse domains, as well as its conceptual simplicity, practical effectiveness, and broad applicability. This is especially evident in healthcare, where most ML models are built on supervised frameworks [11, 21]. Supervised learning excels in tasks with well-labeled data, such as fraud detection [7, 22], image classification [8, 13], and disease diagnoses [11, 21]. Moreover, the transparency of some supervised models—such as decision trees and linear regression [23, 24]–enhances their value in high-stakes fields where interpretability is critical [22, 25]. Within supervised learning, classification tasks are particularly prevalent [13, 18], especially in healthcare domains [11, 12, 21]. Classification involves assigning categorical labels to input data and has diverse applications, from spam filtering and sentiment analysis to facial recognition and disease detection [8,13,26]. Its popularity reflects the intuitive nature of categorical decisions, the interpretability of classical algorithms commonly used (e.g., decision trees [23], logistic regression [24]), and the growing availability of labeled datasets.

Despite its widespread adoption, supervised learning in practice often relies on *black-box* deep

learning models (Fig. 1) that obscure their internal mechanisms and decision-making processes [27–29]. These models, particularly convolutional and transformer-based architectures [30–32], achieve high predictive accuracy but provide little insight into *"why"* a specific decision was made. Their reliance on millions of parameters distributed across complex non-linear transformations makes it difficult to trace outputs back to meaningful input features. As a result, the predictions, while accurate, remain largely opaque to human users [33–35]. This opacity severely limits their utility in real-world settings, particularly as the demand for interpretability grows in domains where machine learning models guide and inform decisions with direct human consequences [33, 36, 37]. In such contexts, individuals have ethical and legal rights to an explanation, ensuring that decisions are transparent and made fairly [38,39]. Without such explanations, many supervised models risk non-compliance in sensitive domains like healthcare or finance.

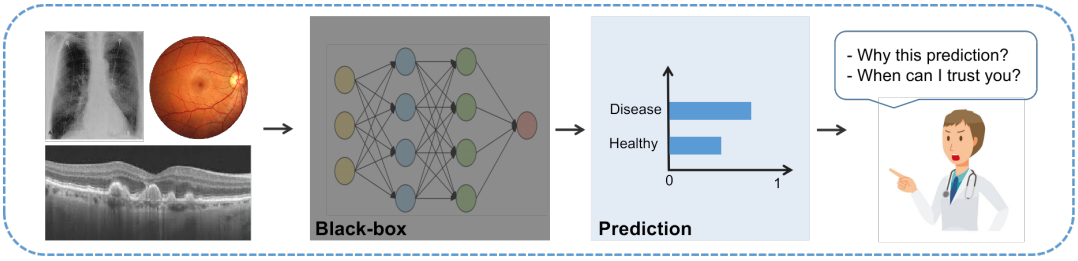

Figure 1: Example of a black-box classifier applied to medical imaging. The black-box model produces a binary classification outcome without offering insight into the underlying decision process, limiting transparency and hindering clinical trust in its predictions.

1.2 Requirements for explainable models

In machine learning, the terms *explainability* and *interpretability* are most often used interchangeably [40,41]. According to *Miller* [41], interpretability is the degree to which a human can understand the cause of a decision. Explainability is essential in machine learning, as it fosters trust, supports accountability, and enables the safe deployment of models in real-world applications. By providing insights into model behavior (Fig. 2), explanations allow users to identify limitations, uncover potential biases, understand underlying data patterns, and recognize opportunities for model improvement or further investigation. The ultimate goal of explanation is to provide sufficient context for humans to assess automated decisions critically—especially to detect and correct potentially harmful or erroneous outcomes [42, 43]. This becomes particularly important when machine learning models are deployed in high-stakes domains like healthcare, where decisions directly impact individuals [36, 38].

Beyond its practical utility, interpretability is increasingly recognized as a legal and ethical imperative. As highlighted by [38,44], the development and deployment of trustworthy and transparent machine learning systems is not merely desirable—it is mandated by law in many jurisdictions [39,44,45]. Recent legislative frameworks, particularly within the European Union [38,39,45], reflect this priority. For instance, the General Data Protection Regulation (GDPR) enshrines a so-called "right to explanation", more accurately described as "right to information" about the logic and consequences of automated decision-making (Articles 13–15) [46]. Building upon the GDPR, the EU Artificial Intelligence Act (AIA) introduces further requirements for interpretability, particularly in *high-risk* domains [45, 47]. These regulations mandate not only technical robustness but also human oversight, demanding that AI systems remain under meaningful human control

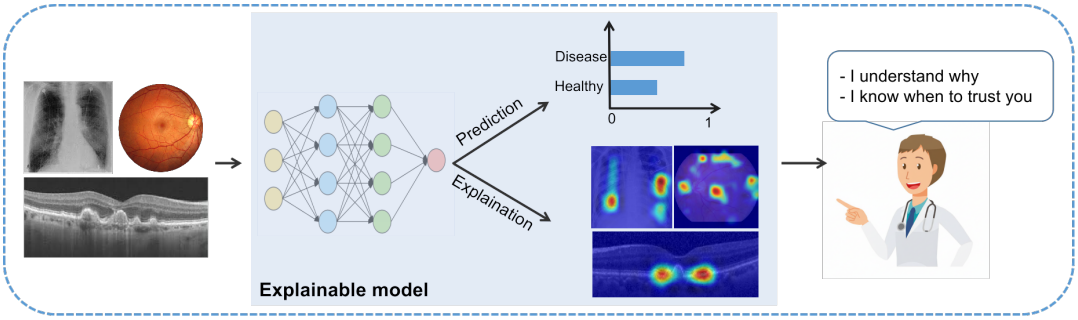

Figure 2: Example of an explainable classifier applied to medical imaging. The model provides interpretable predictions by highlighting the image regions that are most relevant to the predicted outcome, offering insight into the potential decision process and fostering clinical trust. The explanation maps, which indicate region relevance, are overlaid on the corresponding images.

for informed intervention. As *Pavlidis* [48] emphasizes, explanations are "a prerequisite for accountability, fairness, public trust, and effective regulation and supervision".

To meet these regulatory expectations, model interpretability must be approached from both a technically rigorous and user-centric perspective [49–51]. This is essential not only for tasks such as model auditing and validation, but also for ensuring that explanations are meaningful and actionable to diverse stakeholders. Consequently, achieving meaningful interpretability and transparency in machine learning is not solely a technical challenge but a central requirement for the responsible use of AI [39, 45], ensuring that models are accessible, auditable, and aligned with societal and legal expectations.

1.3 Challenges in supervised classifiers

Traditional machine learning models, such as decision trees and logistic regression, are often preferred for their transparency and interpretability. By explicitly linking input features to predictions, these models provide clear reasoning—a key advantage in high-stakes domains such as healthcare and finance [22, 25]. However, their simplicity limits their effectiveness in complex tasks, such as image classification, where the data are high-dimensional and patterns are often hierarchical, making it difficult to capture them through decision trees or linear decision boundaries. To overcome these limitations, deep neural networks have emerged as a powerful alternative, capable of learning meaningful abstract representations from unstructured data [10, 19]. DNNs have achieved state-of-the-art results in numerous vision tasks, including object recognition [8] and medical image classification [14, 21], which has driven their widespread adoption. Despite this success, deep learning models are often criticized for their lack of interpretability [27, 29, 36]. Unlike classical models, deep supervised networks operate largely as black boxes [27, 28, 41], making it difficult to understand how inputs influence outputs—a critical concern in domains that require transparency and accountability.

Beyond their opacity, deep supervised classification models face several practical limitations. They typically require large volumes of labeled training data [52, 53], which can be costly and time-consuming to obtain, especially in medical imaging, where expert annotations are essential. Class imbalance presents another challenge, as rare but critical classes may be underrepresented, leading to degraded model performance [53]. Furthermore, these models often struggle to generalize to out-of-distribution data [54]; for instance, a classifier trained to detect retinal disease in one

population may fail when applied to another due to distribution shifts [54, 55]. Finally, the classification paradigm inherently restricts outputs to predefined categories, limiting the model's ability to capture continuous variation, subtle distinctions, or previously unseen patterns.

To address these limitations, this thesis focuses on improving the transparency and interpretability of supervised deep learning models for medical image classification, aiming to develop models that are not only accurate but also explainable and trustworthy (Fig. 2).

1.4 Contributions

The primary contribution of this thesis lies in the design and evaluation of supervised deep learning classifiers for medical image classification, with a strong emphasis on enhancing model transparency and interpretability. This research addresses several critical challenges in the field, including the limitations of current explainable deep learning models in providing faithful and transparent outputs [56, 57], the lack of quantitative evaluation of self-explainable models in clinical settings, the interpretability–accuracy trade-off inherent in self-explainable architectures [34], and the difficulty of generalizing explainability approaches across different model architectures due to their model-specific design.

The main contributions of this thesis are as follows.

1. The development of **Sparse BagNet** [1], an improved variant of the BagNet model designed to enhance transparency by incorporating a convolutional classifier head and enforcing sparsity in its explanations.

2. A **retrospective clinical study** evaluating the utility of Sparse BagNet to assist ophthalmologists with the detection of early diabetic retinopathy, demonstrating its positive impact on both diagnostic performance and decision-making time [3].

3. The development of **ProtoBagNet** [2], an architecture that combines BagNet's localized receptive field for feature extraction with prototype-based learning, providing both local and global interpretability within a unified framework.

4. The design of a novel protocol, **SoftCAM**, which leverages inherent properties of convolutional operations—such as locality, spatial alignment, and translational invariance—to transform standard black-box CNNs into inherently self-explainable models [5].

5. The generalization of SoftCAM to standard **hybrid CNN-Transformer architectures**, making them fully convolutional and self-explainable [4].

We believe that these contributions will be of significant value to researchers in computer vision, policymakers, and practitioners currently utilizing supervised black-box models in medical imaging. By highlighting the potential benefits of self-explainable deep learning classifiers, this research aims to inform and guide future developments in the field. The findings presented here can support the design of more effective and transparent deep learning models for medical image analysis. Although primarily evaluated in the context of medical imaging, the methods and insights introduced are broadly applicable and may extend to other data modalities and tasks.

In addition to the work presented in this thesis, I have also contributed to several related projects extending the application of the Sparse BagNet model to other clinical tasks [58, 59]. These include its integration into a deep survival model for disease progression risk prediction—specifically for predicting the conversion risk to late-stage age-related macular degeneration (AMD)

from fundus images [58]. The model has also been applied to epiretinal membrane (ERM) and related pathology detection in optical coherence tomography (OCT) images [59]. Furthermore, I have contributed to a federated learning framework that offers interpretability through prototype learning while addressing statistical heterogeneity using lightweight adapter modules that act as compressed surrogates of local models, enabling clients to generalize across diverse data distributions [60].

1.5 Outlines

The remainder of this thesis is structured as follows. **Chapter 2 (Background)** provides a review of explainability in machine learning, introducing key concepts, methods, and challenges. It also briefly discusses applications of machine learning in medical imaging, with emphasis on the disease and imaging modalities considered in this thesis. **Chapter 3 (Publications)** constitutes the core of the work and presents five papers: three published, one accepted for publication, and one available as a preprint. These contributions cover: (1) the development of the Sparse BagNet model, (2) its clinical evaluation for early diabetic retinopathy detection, (3) the ProtoBagNet architecture for local and global explainability, (4) the SoftCAM approach for transforming black-box CNNs into inherently interpretable models, and (5) a self-explainable, fully convolutional hybrid CNN-Transformer architecture. Each paper is summarized with attention to its motivation, methods, key findings, and contributions, followed by a brief discussion. **Chapter 4 (Discussion and conclusion)** synthesizes the main findings and limitations, explores broader implications, and outlines promising directions for future research.

2 Background

This chapter introduces the relevant concepts and background knowledge to understand the context and relevance of this thesis's contributions to explainable machine learning in medical imaging.

2.1 Supervised deep neural network classifiers

Supervised learning, particularly classification, is the most widespread application of deep neural networks (DNNs), thanks to the growing availability of labeled data in various fields and their practical effectiveness across tasks. This is especially true in vision-based domains, such as the diagnosis of medical images [12, 21]. Given a labeled dataset $\mathcal{D} = \{\mathbf{X}_i, y_i\}_{i=1}^N$, where $\mathbf{X}_i \in \mathbb{R}^d$ denotes an input sample and $y_i \in \{1, \ldots, K\}$ the corresponding class label, a supervised DNN classifier learns a parameterized mapping function $\hat{f}_\theta(.) : \mathbb{R}^d \to \mathbb{R}^K$ to approximate the true underlying function f:

$$\hat{\mathbf{y}}_i = \hat{f}_\theta(\mathbf{X}_i) \approx f(\mathbf{X_i}) = \mathbf{y}_i, \tag{1}$$

where θ denotes the model parameters optimized during training. The output value $\hat{y}_i$, known as "logit", represents unnormalized class scores. The predicted label $\bar{y}_i$ corresponds to the class with the highest logit value, defined as:

$$\bar{y}_i = \arg \max_{c \in \{1,\ldots,K\}} \hat{y}_i^c. \tag{2}$$

The learning process involves passing input images through multiple layers that apply successive linear and non-linear transformations, allowing the model to extract increasingly meaningful feature representations. Training is usually performed using gradient-based optimization methods, such as stochastic gradient descent (SGD) [61], in combination with the backpropagation algorithm [62]. The models used in this thesis are mainly feed-forward architectures [56, 63–65], meaning data flows unidirectionally from input to output. We focus on two prominent families: convolutional neural networks (CNNs) [66] and vision transformers (ViTs) [64], both of which have shown strong performance in computer vision and are increasingly applied to medical imaging tasks [32, 66, 67].

2.1.1 Convolutional neural networks

Convolutional neural networks are a class of deep neural networks designed specifically to process image data (Fig. 3). Using convolutional filters, CNNs automatically learn spatial hierarchies of features, making them highly effective for vision tasks, including object recognition and medical image analysis [21, 66]. Their accuracy, scalability, and computational efficiency—particularly with large datasets—have established CNNs as a standard in computer vision. Among the most widely used architectures are ResNet [65] and VGG [63].

VGG. Developed by the Visual Geometry Group at the University of Oxford, the VGG networks [63] achieved widespread recognition after a strong performance in the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) 2014[1] [68]. Their architecture is composed of stacked 3×3 convolutional filters with ReLU activations and max-pooling layers, creating a uniform and straightforward design that allows for deep models with manageable parameter sizes. VGG-16,

[1] `https://www.image-net.org/challenges/LSVRC/2014`

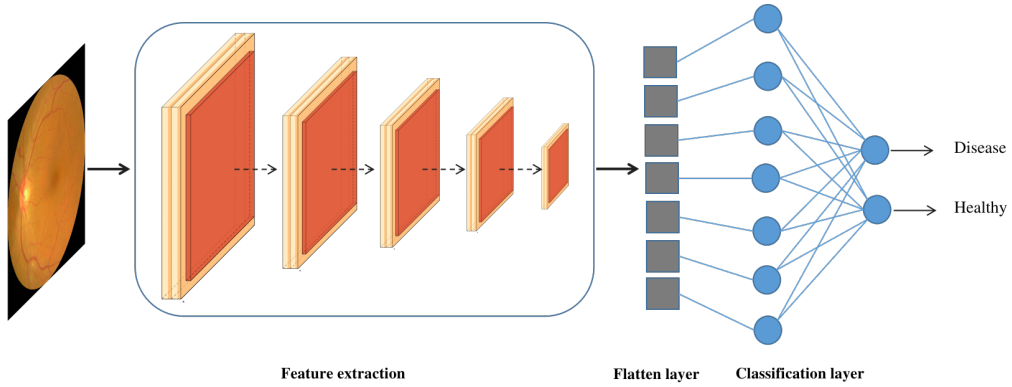

Figure 3: Overview of the Convolutional Neural Network architecture. The input image is processed through a series of convolutional blocks, each composed of convolutional layers followed by pooling layers that progressively extract higher-level and more abstract feature representations. As the spatial resolution decreases across the network, the number of feature maps increases, capturing richer semantic information. The final feature maps are flattened and passed through a fully connected (dense) classification layer that produces the final prediction.

consisting of 13 convolutional layers and 3 fully connected layers, remains widely used despite its high computational cost, particularly as a baseline in transfer learning [66]. However, the very depth that made VGG successful also exposed limitations, such as the *vanishing gradient problem* [69], which later architectures like ResNet explicitly addressed.

ResNet. Deep neural networks are prone to the vanishing gradient problem, in which gradients decrease as they are back-propagated through multiple layers [69]. This hinders effective weight updates, slows convergence, and can make training unstable. Residual Networks (ResNets) mitigate this through residual learning [65]. Instead of learning a direct mapping, each residual block learns a residual function $R(\mathbf{x}_i)$ relative to its input $\mathbf{x}_i$, which is then combined with the input via a skip connection: $\mathbf{x}_i + R(\mathbf{x}_i)$. This design facilitates gradient flow, stabilizes training, and enables the construction of much deeper networks. ResNet gained prominence by winning the ILSVRC challenge in 2015[2] [68]. Among its variants, ResNet-50 offers a strong balance between depth and computational cost, making it a popular choice for classification and detection tasks [66].

2.1.2 Vision transformers

Vision Transformers [64] represent a significant shift in deep learning for visual tasks by adapting the transformer architecture—originally developed for natural language processing [70]—to image analysis (Fig. 4). Unlike CNNs, which rely on convolution to capture local spatial patterns, ViTs partition an image into fixed-size non-overlapping patches and treat each patch as a token, similar to words in a sentence. These tokens are embedded into vectors and processed using self-attention (SA) mechanisms [70], enabling the model to capture long-range dependencies and global context. When trained on large-scale datasets with sufficient computational resources, ViTs achieve performance on par with, and in some cases surpass, CNNs [31, 32]. Prominent architectures include the original Vision Transformer [64] and the Swin Transformer [71].

[2] `https://www.image-net.org/challenges/LSVRC/2015`

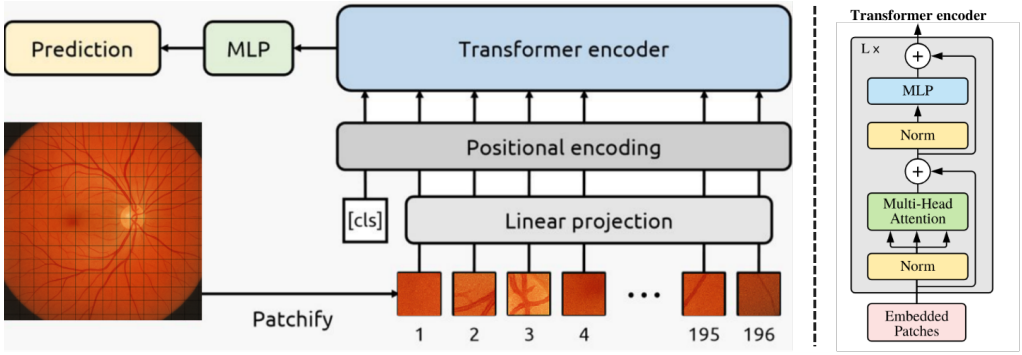

Figure 4: Overview of Vision Transformer models. The input image is split into fixed-size, non-overlapping patches that are linearly projected into embeddings. A learnable classification token [CLS] and positional encodings are then added to these patch embeddings. The resulting token sequence is processed through a stack of transformer encoder layers, each consisting of multi-head self-attention and feed-forward blocks with normalization and residual connections. Finally, the output corresponding to the [CLS] token from the last encoder layer is passed through a multi-layer perception (MLP) head to generate the final prediction. This figure was inspired by Dosovitskiy et al. [64].

Standard Vision Transformer. The standard vision transformer [64] applies the standard transformer encoder architecture with self-attention [70] directly to the image patches. An input image is divided into non-overlapping patches (e.g., 16×16 pixels), which are flattened and linearly projected into embedding vectors. Positional encodings are added to preserve spatial information, and the resulting sequence is processed by transformer layers using multi-head self-attention (MHSA). This global attention mechanism enables ViT to effectively capture long-range dependencies across image regions. Although ViTs achieve strong performance, they generally require large-scale pretraining and substantial computational resources to match CNNs [31, 32].

Swin Transformer. The Swin Transformer (Shifted Window Transformer) [71] was proposed to overcome key limitations of vision transformers by introducing a hierarchical and computationally efficient architecture. Instead of applying global self-attention across the entire image, Swin partitions feature maps into non-overlapping windows and performs local self-attention within each window, greatly reducing computational cost. To allow information exchange across windows, the windows are cyclically shifted in subsequent layers, enabling global context modeling through Shifted Window Multi-Head Self-Attention (SW-MHSA). Furthermore, Swin adopts a multiscale representation strategy similar to that of CNNs, capturing fine-to-coarse spatial hierarchies that are crucial for dense prediction tasks such as segmentation. Thanks to these scalable and flexible designs, the Swin Transformer has demonstrated state-of-the-art performance in various vision tasks, including image classification, object detection, and semantic segmentation [30, 67].

2.2 Machine learning for medical imaging

Machine learning has become an essential component in medical imaging, where accurate and efficient analysis is crucial for disease diagnosis and clinical decision-making. Among various paradigms, supervised learning—especially classification tasks [11, 21]—is the most prevalent, leveraging labeled datasets to train models capable of detecting and categorizing pathological patterns. Such models have reached and, in some cases, exceeded human expert performance [6, 12], particularly in areas such as radiology [17, 72, 73] and ophthalmology [14, 15, 74].

2.2.1 Machine learning in radiology and ophthalmology

Radiology was the first medical specialty to integrate artificial intelligence into disease diagnosis [75, 76], with early work in the 1970s applying computer-aided detection and pattern recognition techniques on chest X-rays (CXR) [76]. Today, it remains the leading field for regulatory-approved AI devices [12,37,77], reflecting its central role in the clinical adoption of machine learning models. Among radiological modalities, CXR (Fig. 5a) is the most widely used, with more than 2 billion scans performed globally each year [77]. Its ubiquity, low cost, and high diagnostic value for common conditions such as pneumonia, tuberculosis, and lung cancer make it a prime target for AI applications [17, 72, 78]. Supervised learning models trained on large-scale CXR datasets now deliver strong performance in automated detection, risk stratification, and clinical decision support, positioning them at the core of radiology-focused machine learning research.

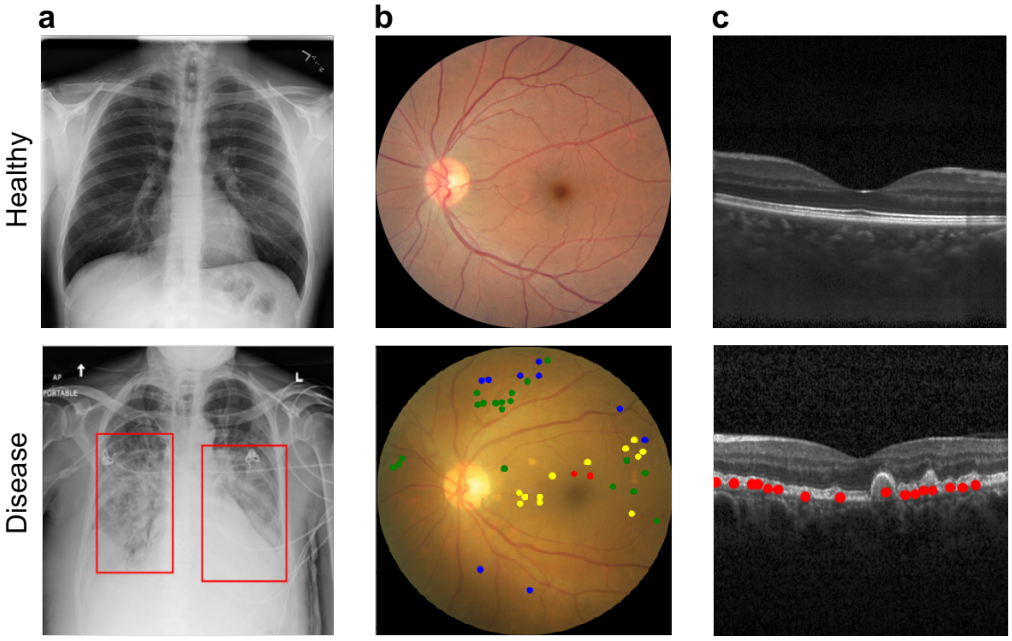

Figure 5: Example of image modalities in radiology and ophthalmology. The first column (**a**) shows chest X-ray images from the RSNA dataset [79], the second column (**b**) presents color fundus photographs from the Kaggle dataset [80], and the third column (**c**) displays retinal OCT images from the Kermany dataset [81]. The top row depicts healthy cases, while the bottom row illustrates disease examples. In the chest X-ray, red bounding boxes highlight pneumonia-related opacities. The diseased fundus image includes annotated retinal lesions characteristic of diabetic retinopathy, where red points denote microaneurysms, blue indicate hemorrhages, yellow represent soft exudates, and green correspond to hard exudates. Red markers on the OCT image identify drusen lesion deposits typical of AMD.

In parallel, ophthalmology has rapidly emerged as another major domain for medical AI [12,77]. The early and widespread digitization of retinal imaging [82], together with large-scale screening programs such as EyePACS [83] facilitated the development of deep learning systems that achieved impressive accuracy in eye disease detection [14, 15, 74], placing ophthalmology among the leading specialties with multiple regulatory-approved AI medical devices [12, 49]. Ophthalmologists primarily rely on two imaging modalities: color fundus photography (Fig. 5b) and optical coherence tomography (OCT) (Fig. 5c). Fundus imaging provides two-dimensional views of the retina, capturing key structures such as the optic disc, macula, and blood vessels. In contrast, OCT produces high-resolution cross-sectional scans of retinal layers, enabling detailed structural as-

sessment. Together, these modalities are used for the diagnosis of retinal diseases [81, 82, 84, 85], including diabetic retinopathy (DR), age-related macular degeneration (AMD), diabetic macular edema (DME), and epiretinal membrane (ERM).

Beyond ocular diseases, retinal images can also reveal systemic health conditions [86], providing insights into cardiovascular disease, diabetes, hypertension, and even neurodegenerative disorders. Similarly, chest X-ray imaging, traditionally used for pulmonary diagnoses, has been shown to capture systemic information [87,88], including cardiovascular disease, hypertension and vascular health, metabolic disorders, and even biological age and mortality risk. This broader diagnostic potential makes both ophthalmology and radiology rich domains for supervised machine learning, enabling the detection of localized pathology, systemic health markers, and even protected attributes such as gender predictions [89, 90].

In this study, we focus specifically on using retinal images and chest radiographs for the detection of lesions and pathology, as well as the classification of diseases within their respective imaging domains.

2.2.2 Challenges of machine learning classifiers

Despite their widespread success in computer vision, deep neural networks face several key challenges that can compromise their reliability and hinder their adoption in real-world applications.

- **Data requirements.** DNN classifiers, particularly ViTs and CNNs, typically require large amounts of labeled data to achieve good generalization and avoid overfitting [9,31]. In fields like medical imaging, the acquisition of such datasets is costly and time-intensive due to the need for expert annotations [52].

- **Computational cost.** Training and deploying deep learning models often require substantial computational resources, including high-performance GPUs [9, 32]. Consequently, their deployment is less feasible in resource-constrained environments such as mobile devices, edge platforms, or real-time applications.

- **Robustness.** DNNs are highly sensitive to small perturbations in input data [91, 92], such as noise, adversarial attacks, and distribution shifts (e.g., changes in lighting, background, or acquisition devices) [9, 55]. These vulnerabilities can substantially degrade model performance in real-world settings, where data is rarely independently and identically distributed.

- **Interpretability.** Most DNNs are black boxes that provide little insight into their decision-making processes [27, 28]. In healthcare, this opacity hinders clinicians' ability to verify or trust the model outputs, compromising safety and accountability. The lack of transparency also raises ethical and legal concerns, particularly regarding compliance with the EU Artificial Intelligence Act [45] and the General Data Protection Regulation (GDPR) [39, 46].

- **Bias and fairness.** Models trained on unbalanced or biased datasets can propagate and even amplify these biases, resulting in unfair and unreliable outcomes for underrepresented groups [93, 94]. This issue persists despite the use of large-scale datasets, which often fail to capture the full diversity of real-world data.

- **Overfitting.** Deep models are prone to overfitting [9], particularly when trained on small, imbalanced, or low-diversity datasets [52, 53]. Without appropriate mitigation strategies,

they may memorize the training samples instead of learning patterns that are generalized to unseen examples.

Addressing these challenges remains a central research focus of ongoing research. Current strategies include data-efficient learning [95, 96], model compression [95, 97], adversarial training [91, 92], fairness-based training [93, 94], and interpretability methods [25, 29, 36]. This thesis aims primarily to enhance model interpretability by developing self-explainable supervised deep learning classifiers for medical image diagnosis. In parallel, it addresses data and computational constraints through the design of hybrid CNN-Transformer models that are both efficient and inherently interpretable.

2.3 Explainability of DNN classifiers

As machine learning models become increasingly complex, understanding their decision-making processes has become crucial, particularly in sensitive domains that require transparency, accountability, and trust. This section introduces the foundations of Explainable AI (XAI), highlighting its motivations, key concepts, and commonly used terminology.

2.3.1 Motivations for interpretable AI models

The rise of deep learning has introduced powerful, yet opaque, models. Although mathematically well-defined, DNNs are highly non-linear and complex, often described as "black boxes" [27, 28, 48]. This lack of transparency raises concerns in high-stakes fields where trust, accountability, and ethical responsibility are essential [38, 49, 51]. Explainable AI aims to address these issues by making model decisions more transparent and interpretable. Instead of simply producing predictions, explainable models provide human-understandable justifications through language [98, 99], visualizations [28, 29], or logical reasoning [23, 24]. Effective explanations require not only interpretable outputs but also insight into the model's internal reasoning.

The prevalence of black-box models comes from the test set paradigm [100], which prioritizes predictive accuracy on held-out data over interpretability—ultimately fostering the development of increasingly opaque systems. In sensitive applications such as medical diagnosis, this opacity is particularly concerning, as unexplained errors or biases can have serious consequences [45, 46]. Without interpretable outputs, the detection of bias or validating decisions becomes challenging, hindering the deployment of trustworthy AI [25, 49, 94]. In response, explainable AI has emerged as a critical research area, driven by interpretability and transparency, scientific discovery, and regulatory compliance.

Interpretability and transparency. As machine learning systems increasingly inform decisions across diverse domains, the need for interpretability and transparency has become paramount. Although black-box models often achieve strong predictive performance, their opacity can undermine human trust and hinder adoption—particularly in applications that directly impact individuals [36, 38]. Explainability seeks to address this challenge by clarifying *how* and *why* models make their predictions, providing insight into their underlying reasoning. Moreover, it facilitates a more effective human–AI interaction [50, 74, 101], allowing users to interpret better and evaluate model outputs. In clinical settings, interpretability not only fosters trust [102] but also strengthens collaboration between AI systems and healthcare professionals [74, 101], ultimately supporting a safer and more ethical deployment.

Scientific discovery. Beyond interpretability and transparency, scientific discovery requires not only accurate predictions but also a clear understanding of how those predictions are made. Researchers aim to validate hypotheses and uncover new insights, yet black-box models often hinder this process, limiting their value in discovery-driven research. Explainable AI can address this gap by offering tools that reveal how models make decisions, providing interpretable insights while maintaining strong predictive performance. For example, feature attribution methods have allowed discoveries in cancer pharmacology [103]. More broadly, explainable AI supports critical tasks such as model debugging [42, 104], bias detection [94, 104], and error analysis [43, 104], improving the model development pipeline and fostering new avenues for scientific findings.

Regulatory compliance. Legal and ethical standards are increasingly shaping research on explainable AI. Regulatory frameworks such as the EU GDPR [39, 46] grant individuals the right to meaningful information on the logic involved in automated decision-making (Articles 13–15) and limit the use of automated decisions that significantly affect individuals (Article 22). Similar requirements are emerging around the world, including the US and the UK [44]. These regulations provide a strong compliance-driven incentive for organizations to develop AI systems that are not only accurate but also explainable, transparent, and aligned with ethical and legal standards.

Taken together, these motivations highlight that explainability is no longer optional but a fundamental requirement for the responsible and trustworthy deployment of AI systems.

2.3.2 Key concepts in explainable AI

Explainable and interpretable AI, along with the related concepts of "explainability" and "interpretability", have gained significant attention in the machine learning community. Although often used interchangeably [40], these terms generally refer to approaches that enable humans to understand the rationale behind a model's decisions [40, 41]. While there is no universally accepted definition, explainability is commonly described as the ability to describe, in human-understandable terms, how a model processes input data to produce learned representations and outputs. Despite their widespread use, these concepts remain imprecisely defined and are subject to ongoing debate in the academic literature [29, 40, 41, 105]. In this thesis, the terms are used interchangeably to denote the degree to which a human can understand the internal reasoning or decision process of a machine learning model [41].

The interpretability of machine learning models can be characterized along three primary dimensions [25, 27, 28, 40] (Fig. 6): (i) intrinsic vs. post-hoc, (ii) global vs. local, and (iii) model-specific vs. model-agnostic.

Intrinsic vs. Post-hoc Interpretability. This dimension differentiates models that are inherently interpretable from those that require external explanations after training. *Intrinsically interpretable* (or *interpretable-by-design*) models integrate transparency into their architecture [34, 106], often through constrained or modular designs that make their decision-making process explicit [56, 57, 107]. Such models are often referred to as *self-explainable*, as interpretability is an intentional design feature rather than post-hoc. In contrast, post-hoc interpretability methods aim to explain already-trained black-box models [27, 28] using techniques such as saliency or feature attribution maps [108–110], which highlight discriminative image regions to a model prediction. However, inherently interpretable models often trade off predictive performance [34].

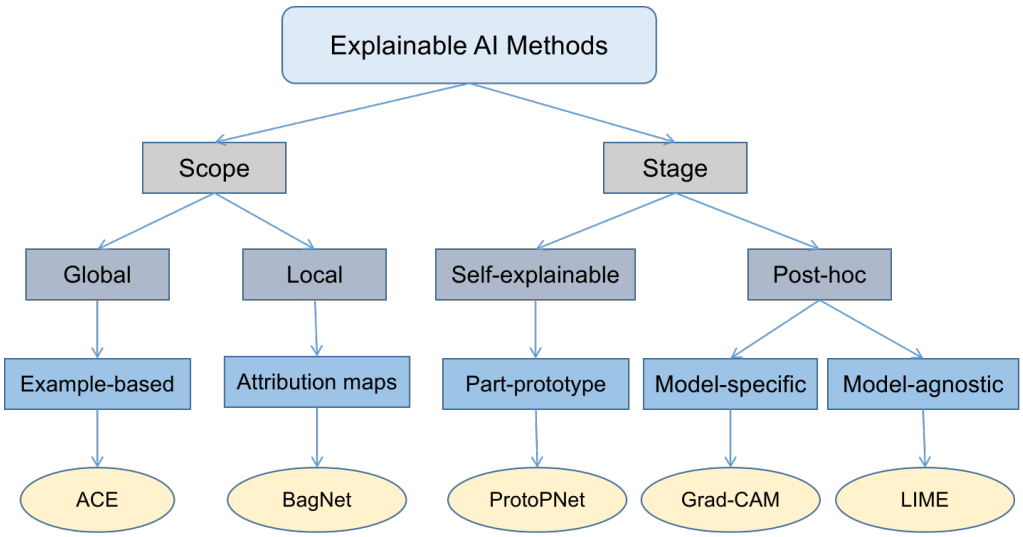

Figure 6: Types of interpretability. Overview of XAI methods, including post-hoc and self-explainable approaches. Leaf nodes provide specific examples corresponding to each category.

Global vs. Local Explanation. The scope of interpretability can be either *local* or *global*. Local explanations clarify the rationale behind a model's decision for specific input instances, for example, by highlighting the salient image regions most relevant to a classification outcome [111,112]. Global explanations, in contrast, characterize the general behavior of a model throughout the entire dataset, identifying features that are consistently important or describing decision boundaries [57,113]. While local interpretability aids in understanding and validating individual predictions, global interpretability provides a broader perspective on the model's overall logic.

Model-specific vs. Model-agnostic Methods. Explainability techniques can also be categorized based on their reliance on a model's internal structure. *Model-specific* methods are tailored to particular model families and require access to internal components such as parameters, activations, or gradients. By leveraging this internal knowledge, they can produce precise and efficient explanations aligned with the model's architecture. Examples include Grad-CAM for CNNs [108], attention maps for Vision Transformers (ViTs) [67], and built-in feature importance scores for decision trees [110]. In contrast, *model-agnostic* methods treat the model as a black box, relying solely on its inputs and outputs to infer which features influence predictions. Methods such as LIME [114] and SHAP [115] illustrate this category, offering broad applicability across diverse architectures. However, this flexibility often comes at the expense of computational efficiency and explanation precision compared to model-specific approaches, which are often derived from self-explainable models [56, 57].

In summary, deep learning models often operate as "black boxes", raising concerns about trust, accountability, and reliability in high-stakes domains. Explainable AI addresses these issues by promoting model transparency, supporting scientific discovery, and facilitating regulatory compliance. Interpretability can be achieved either intrinsically—built into the model—or post-hoc, derived after training. In this thesis, the terms "self-explainable" and "interpretable-by-design" are used interchangeably to emphasize that interpretability is an intentional property arising from a model's architecture and learning process. The next section introduces commonly used post-hoc explanation techniques for black-box models, alongside self-explainable model architectures.

2.4 Post-hoc explanations

Post-hoc interpretability methods aim to explain the predictions of already-trained black-box models without modifying their internal structure. These approaches are particularly suitable for complex black-box models such as CNNs and are widely adopted due to their flexibility, which does not affect model performance [27,28,116]. Most post-hoc methods provide local explanations, focusing on "why" a particular prediction was made, although a few also offer global insight into the model's behavior throughout the entire dataset. A prominent category of post-hoc techniques is feature attribution maps (Fig. 7), which assign importance scores to the input features based on their potential contribution to the output. These techniques span various data modalities, including scalar attributions for tabular data [114,115] and saliency maps for images [108,111,117].

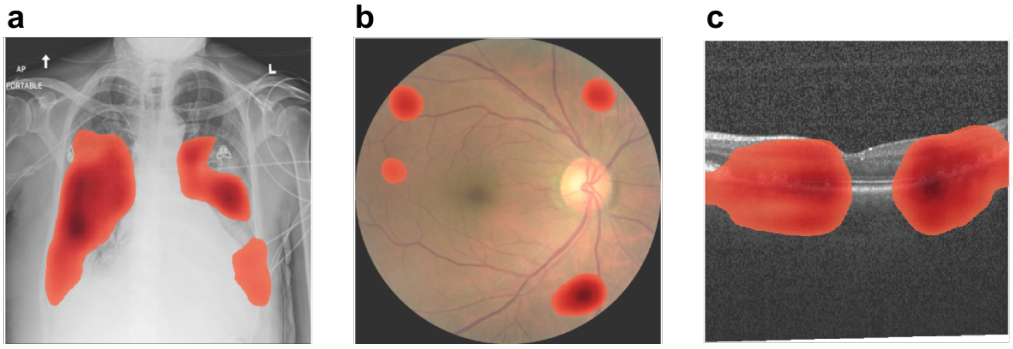

Figure 7: Example of Grad-CAM attribution maps from a VGG-16 model on disease medical imaging. The VGG-16 model was trained for pneumonia detection on chest X-ray images (**a**), diabetic retinopathy detection on fundus images (**b**), and drusen detection on retinal OCT scans (**c**). Grad-CAM was then applied post-hoc to generate attribution maps for disease-positive images, highlighting in red the regions strongly associated with pathological features or lesions. The resulting maps were upsampled and overlaid on the corresponding images, providing visual explanations of the model's predictions.

Post-hoc methods can be broadly grouped into several major categories: feature attribution methods [108, 111], concept-based explanations [118–120], and example-based reasoning [28, 121].

2.4.1 Feature attribution methods

Feature attribution—also known as feature importance—methods quantify how much each input feature contributes to a model's prediction (Fig. 7). They are typically grouped into gradient-free, gradient-based, and perturbation-based approaches [114–116, 122, 123]. Gradient-free methods [111, 112] rely on specific CNN architectures, such as models with a global average pooling (GAP) layer before the final fully connected (classification) layer. They compute class-specific attribution by taking a weighted sum of the final convolutional feature maps, where the learned class weights act as importance scores. Gradient-based methods, such as Grad-CAM [108] and Layer-CAM [109], instead use the gradient of the class score with respect to convolutional feature maps as importance weights, producing heatmaps that highlight the regions most relevant to the model's prediction. However, raw gradients can be noisy and unstable [112], which has motivated the development of improved techniques such as Integrated Gradients [117], Guided Backpropagation [124], and Layer Relevance Propagation [125]. These methods improve the faithfulness, smoothness, and human interpretability of gradient-based explanations. Perturbation-based methods determine importance by systematically modifying or masking part of the input and observing changes in the model's output [126]. Being model-agnostic, they can explain any black-box model.

Prominent examples include LIME (Local Interpretable Model-agnostic Explanations) [114], which builds local surrogate models; SHAP (Shapley Additive Explanations) [115], which uses Shapley values to estimate feature contributions; and RISE (Randomized Input Sampling for Explanation) [116], which randomly perturbs image regions to assess their influence on predictions. While intuitive and widely applicable, perturbation-based methods are often computationally intensive and less scalable for high-dimensional inputs [126].

2.4.2 Concept-based explanations

Concept-based methods explain predictions through high-level human-understandable concepts rather than low-level input features [118], bridging the gap between learning representations and domain knowledge. A concept can be defined as a semantic unit of meaning that connects human understanding to model representations—such as edges, color, texture, or lesion boundary in medical imaging—and serves as an interpretable building block that links low-level model features to high-level human reasoning. Concepts are typically applied post-hoc, and can be manually defined, automatically discovered, or generated through a combination of both approaches [127,128]. Concept-based methods assess whether a model's decisions are driven by meaningful semantic attributes or other domain-relevant features, rather than raw input patterns. Prominent approaches include TCAV (Testing with Concept Activation Vectors) [120], which quantifies the sensitivity of internal activations to predefined concepts by averaging activations across concept examples to form concept activation vectors that directly link model behavior to expert knowledge; another well-known example is ACE (Automated Concept-based Explanations) [119], which automatically discovers interpretable concepts by clustering similar image patches without manual labeling and then applies TCAV to quantify their influence; and ConceptSHAP [128], which extends SHAP to the concept-level, providing explanations in terms of human-meaningful attributes rather than raw features. By grounding explanations in semantically meaningful terms, concept-based methods enhance post-hoc interpretability, making model reasoning more intuitive, transparent, and aligned with human understanding.

2.4.3 Example-based explanations

In contrast to concept-based approaches that explain model decisions through high-level semantic concepts, example-based explanations justify predictions using specific data instances, such as influential or similar training samples [121, 129]. These methods provide concrete, case-based reasoning by identifying examples that the model referenced or learned from when making a prediction, offering an interpretable rationale in the form of "*the model predicted this because it resembles these examples*". This approach shifts the focus from abstract feature importance to concrete evidence, helping users understand decisions through relatable examples [129,130]. Common techniques include nearest neighbor (KNN) methods [121], which retrieve training instances most similar to a given input in the model's embedding space; prototype and criticism [130], which search representative examples (prototypes) for each class and identify outliers where the model performs poorly; and influence functions [129], which estimate how individual training samples affect specific predictions, thus providing traceability between training data and model outcomes.

Summary. Post-hoc explanation methods are widely used to interpret supervised black-box models. Among them, feature attribution–based methods are the most commonly used, as they in-

dicate "*where*" the model focuses by highlighting discriminative regions or features, while concept-based and example-based approaches explain "*why*" a decision is made by linking predictions to meaningful concepts or influential training examples. Another post-hoc approach is counterfactual explanations [90,90,131], which identify the minimal input changes needed to alter a prediction but can be computationally expensive, particularly in high-dimensional spaces. However, post-hoc explanations often approximate rather than faithfully reflect model reasoning and can be sensitive to factors like learning rate or initializations [43, 132], leading to instability and inconsistent attributions [43, 132]. In contrast, self-explainable models provide interpretability by design [33, 34, 106], generating predictions and explanations within a unified framework, which mitigates approximation errors and improves trustworthiness in high-stakes decision-making [33, 38, 51].

2.5 Self-explainable models

Inherently interpretable models embed explanations directly into their design, eliminating the need for external post-hoc methods. This is achieved through architectural constraints or training objectives that promote transparency [33, 34, 106], although such design choices sometimes come at the cost of predictive performance. Classical interpretable models, such as linear regressors and decision trees [23, 24], are straightforward to understand but lack the capacity and flexibility of deep neural networks—which are universal function approximators [133]—making them less suitable for complex tasks like image analysis. To address this limitation, recent research has developed self-explainable deep learning architectures that aim to balance interpretability and predictive performance [34]. In computer vision, promising approaches include part-prototype networks [57, 134] and bag-of-local-features models [56, 135].

2.5.1 Part-prototype networks

Part-prototype networks (Fig. 8) are a class of interpretable-by-design deep learning models that explain their predictions by comparing parts of an input (e.g., image patches) to a set of learned prototypes—representative patterns derived from the training data that correspond to meaningful local concepts—following the principle of "*this looks like that*" [57]. Each prototype captures a semantically relevant visual feature, such as a lesion type like *drusen*, *microaneurysms*, or *hemorrhages* in retinal disease classification. The model then predicts by quantifying how similar regions of a new input are to these learned prototypes, producing similarity maps that highlight discriminative regions. This framework enhances transparency by allowing users to inspect which prototypes guided a decision. Conceptually, prototype networks are closely related to concept-based explanations, as both aim to provide human-interpretable reasoning; however, part-prototype models automatically learn visual concepts from the latent feature space during training [134, 136], rather than relying on predefined semantic concepts as in Concept Bottleneck Models [107]. Prototypes can be either explicit—directly corresponding to real image patches—or implicit [137], learned from the latent space without direct correspondence to input images or patches.

A foundational explicit part-prototype model, ProtoPNet [57], learns class-specific prototypes and links them to the most similar patches from the training data for interpretability. Several extensions refine this idea: ProtoCAPs [138] combines prototype learning with capsule networks to leverage privileged information; MProtoNet [139] adapts ProtoPNet for 3D multi-parametric MRI brain tumor classification; and XProtoNet [16] introduces spatial flexibility to improve interpretability in chest radiography. By grounding predictions in visual examples, part-prototype

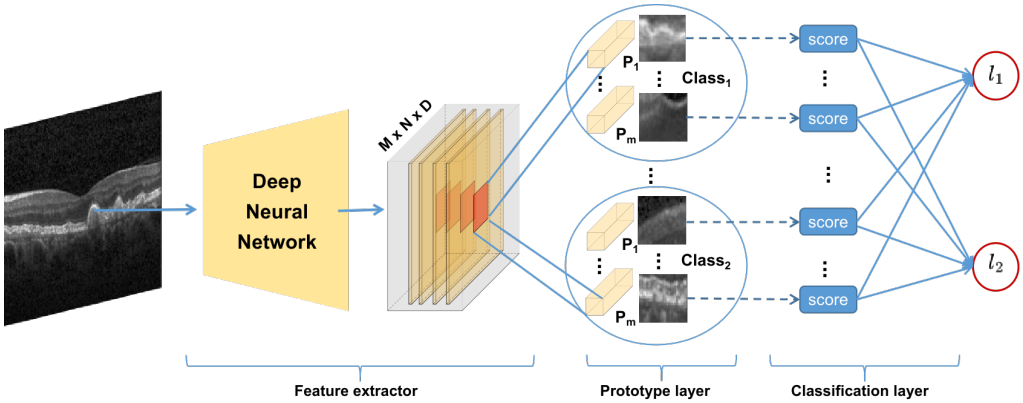

Figure 8: Overview of part-prototype networks. A typical part-prototype model comprise in three main components: (i) a feature extractor—commonly a CNN—that encodes input images into deep feature representations; (ii) a prototype layer that learns a set of class-specific prototypes by optimizing feature embeddings to represent interpretable parts, such that prototype closely matches real image patches from its class while remaining dissimilar to others; and (iii) a classification layer that linearly combines the similarity scores between input features and class-specific prototypes so that classes and class-specific prototypes, assigning highest logits to classes with the most strongly activated prototypes. This design enables simultaneous classification and interpretation. This figure was inspired by Djoumessi et al. [2]

networks provide intuitive and transparent decision reasoning, although challenges such as prototype redundancy [140] and spatial imprecision in explanations [141, 142] persist.

2.5.2 Bag-of-local-features models

Bag-of-local-feature models make predictions by analyzing small, localized regions of an input rather than relying on global representations [135]. This localized design enhances interpretability by allowing decisions to be directly traced to specific image patches—a valuable property for fine-grained visual tasks where discriminative features are spatially distributed, such as in diabetic retinopathy detection [84, 143]. A leading model in this category is BagNet [56], which modifies the ResNet-50 architecture [65] to restrict its receptive field to small patches (e.g., 9×9, 17×17, or 33×33 pixels). It achieves this by replacing most 3×3 convolutions with 1×1 filters and reducing the stride, producing an implicit grid of high-dimensional, patch-level features in the penultimate layer (Fig. 9). These local features are then average-pooled and passed through a linear classifier to obtain the final prediction. Because the final operations are linear, BagNet can compute class logits independently for each patch, allowing patch-wise contribution to be visualized directly as heatmaps—providing built-in interpretability without post-hoc explanations. Conceptually, BagNet extends the traditional bag-of-visual-words paradigm [135], treating an image as a collection of local patches whose individual scores are averaged to form the overall prediction. Despite its limited receptive field, BagNet performs competitively on challenging vision tasks [144–146], illustrating that interpretability can be achieved without a substantial loss in accuracy.

However, the model's focus on local features limits its ability to capture global context and long-range dependencies [147], which can hinder performance in tasks that require global understanding or larger receptive fields. Additionally, performing average pooling on the penultimate feature map followed by a fully connected classification layer discards spatial relationships (Fig. 9), which could otherwise improve transparency, and sometimes necessitates explicit patch splitting with extra forward passes to generate explanations. Nonetheless, the model's reliance on patch-

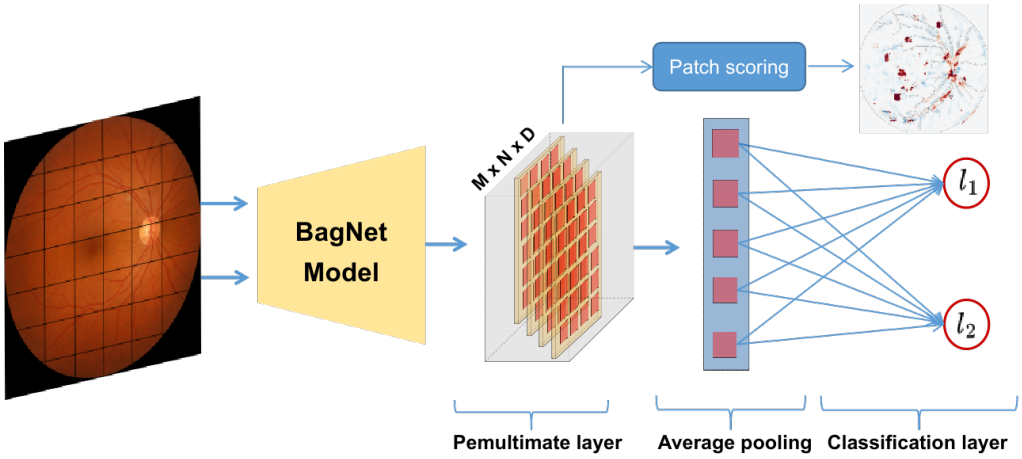

Figure 9: Overview of the BagNet architecture. During training, the model extracts features from small image patches, which are averaged across the channel dimension via global pooling and then passed to a linear classifier to predict class probabilities. During inference, the image can be explicitly split into patches and processed individually, producing one logit per class that represents local evidence. These logits can be spatially combined and visualized as class-specific saliency maps, providing explanations of the model's prediction.

level aggregation, architectural simplicity, and strong empirical performance make bag-of-local-feature models a compelling choice for applications where localized evidence and interpretability are essential for decision-making.

Summary. Self-explainable models provide inherent interpretability, offering a strong alternative to post-hoc explanation methods. Prototype-based models ground decisions in representative example patches but can suffer from redundancy and coarse explanations [140, 142], while bag-of-local-features models enable patch-level interpretability via feature map attributions but may miss global context crucial for complex tasks [147]. Another family, dynamic alignment networks [148, 149], offer fully decomposable, theoretically faithful explanations by aligning the learned weights with the input features during training, although they remain computationally expensive and less effective in capturing nonlinear interactions. This thesis primarily builds upon bag-of-local-features models and part-prototype models. More broadly, explainable models are often tied to specific design choices, limiting their generalizability to other architectures. Moreover, the diversity of interpretability paradigms—from part-prototypes to local-feature models and post-hoc methods—makes evaluation challenging, as no single metric can comprehensively capture their varied explainability objectives.

2.6 Evaluating explanations

The spectrum of explainable AI techniques—ranging from post-hoc methods to self-explainable models—demonstrates that explanations can be provided in fundamentally different ways. Post-hoc methods, such as counterfactuals, attribution maps, and example-based explanations, provide interpretability after model training, often approximating rather than faithfully revealing the model's internal reasoning [27,28]. In contrast, self-explainable models—including prototype-based architectures, and bag-of-local-features models—embed interpretability into their architecture, enabling explanations that are inherently tied to the model's computations [34, 106].

Evaluating explanations across diverse approaches—particularly in medical imaging—remains a major challenge [150, 151]. A central difficulty is the lack of a universally accepted ground truth for what constitutes a "correct" explanation. Furthermore, differences in model architecture and explainability techniques hinder the establishment of standardized and comparable evaluation frameworks. Current evaluation strategies are commonly grouped into two categories: [150–152]: (1) *human-centered evaluations*, which rely on expert or user judgment to assess the clarity, trustworthiness, and clinical relevance of explanations in real-world contexts, and (2) *functionality-grounded evaluations*, which use automated, quantitative metrics to measure explanation properties—such as faithfulness or localization—independent of human assessment.

2.6.1 Human-centered evaluation

Human-centered evaluations assess the quality of explanations from the perspective of end-users—especially domain experts—to determine whether they are understandable, useful, trustworthy, and clinically meaningful (especially in medical contexts). Qualitative approaches include *human interpretability* [150, 152, 153], which measures how easily users understand explanations through user studies, structured interviews, or questionnaires; *usefulness or actionability* [150, 153], which evaluates whether explanations assist users in making better or more informed decisions (e.g., identifying model errors or verifying predictions); *trust and satisfaction* [150, 153], which assess whether explanations increase users' confidence in the model's predictions; *plausibility* [150, 153], which examines whether explanations appear reasonable and consistent with human intuition; *domain expert validation* [150, 152, 153], which checks alignment with expert knowledge and clinically relevant features; and *faithfulness perception* [154], which captures whether users believe explanations truly reflect the model's reasoning.

While these assessments are inherently qualitative, they can be quantified through structured methodologies such as rating scales, task-based performance metrics (comparing accuracy, speed, or confidence with and without explanations), expert agreement measures, and survey-based aggregation, which translate open-ended feedback into numerical or frequency-based summaries [153].

2.6.2 Functionality-grounded evaluation

Functionality-based approaches automate evaluations without requiring direct human input, making them particularly valuable for benchmarking large numbers of models or methods efficiently. These approaches are widely used to evaluate both post-hoc and self-explainable models, quantitatively assessing whether the explanations faithfully reflect true decision-making process of the model and are consistent with expert or domain knowledge. Two key quantitative metrics frequently used *fidelity* (or *faithfulness*) [116, 153, 155] and *localization precision* [73, 156], both assessing how accurately an explanation reflects the model's true reasoning or aligns with domain knowledge. A typical approach involves *feature occlusion*, in which features identified as important are removed to observe the resulting drop in model confidence [126]. In contrast, the complementary *insertion metric* progressively adds features back in order of their importance, measuring the improvement in prediction confidence [116]. Localization precision, on the other hand, evaluates the spatial precision of attribution maps by comparing them with ground truth annotations to determine whether class-relevant regions are correctly identified [73, 156]. However, obtaining detailed annotations can be expensive and time-consuming, especially in medical domains.

Although functionality-based metrics offer objectivity, reproducibility, and scalability, they

primarily evaluate how well explanations align with model behavior rather than human reasoning, indirectly capturing domain alignment through the use of ground truth labels.

Summary. Explanations are crucial for both post-hoc and self-explainable models, ensuring that interpretability aligns with expert knowledge and clinical reasoning [26, 35]. Effective XAI methods should balance qualitative and quantitative evaluations [126, 152], although standardization is challenging due to the architectural diversity of approaches. Post-hoc methods require fidelity assessments to verify that explanations faithfully reflect model reasoning, whereas self-explainable architectures must demonstrate that their built-in interpretability faithfully represents their computations. Evaluation criteria also vary by paradigm: concept-based methods are assessed by the clarity and relevance of learned concepts [118]; prototype-based models by the quality and diversity of prototypes [157], and example-based methods by the stability and consistency of explanations across similar inputs [158]. Human-centered evaluations offer practical insight, but are inherently subjective and resource-intensive, while functionality-grounded metrics offer objectivity and scalability, but may overlook human perspectives. Ultimately, the choice of the evaluation approach should align with the application: trust-critical domains benefit from human-based studies, whereas large-scale benchmarking favors automated assessments. Combining qualitative and quantitative metrics ensures that explanations are both faithful to the model and meaningful and actionable for end users. Developing standardized, architecture-agnostic frameworks that integrate both perspectives remains a key research challenge.

2.7 Advancing self-explainable AI

Explainable AI has become a crucial research area in computer vision, driven by the need to make complex models more transparent and understandable to human users [36]. Psychological studies emphasize that explanations are fundamental to human reasoning and learning, enabling users to understand how predictions are formed and to integrate new information with previous knowledge [41, 51, 159]. This reinforces the growing importance of XAI—particularly self-explainable models—in medical image analysis, where interpretability and transparency are essential for clinical adoption. However, despite their advantages over post-hoc approaches, self-explainable models face several challenges including slower development due to design constraints, inherent trade-offs between accuracy and interpretability, overstated claims of explainability, and limited generalizability across architectures like CNNs and ViTs.

Building on this foundation, this thesis advances the field of self-explainable AI by enhancing the interpretability and transparency of explainable models, with a particular focus on BagNet and prototype-based architectures. We improve BagNet's transparency and demonstrate its clinical relevance by showing that its explanations can improve clinician performance. Furthermore, we introduce a unified interpretability framework that integrates BagNet's localized receptive field with part-prototype networks, while proposing a systematic protocol to transform black-box models into self-explainable systems. The resulting model and explanations are rigorously evaluated using both human-centered and functionality-grounded metrics to ensure faithfulness, usability, and clinical alignment.

The following chapter elaborates on these contributions, with the methodology, experimental, and key results summarized.

3 Publications

Based on the interpretability challenges outlined in the previous chapter, this chapter summarizes the four first-author peer-reviewed papers and one preprint that form the core contributions of this thesis. The peer-reviewed works include: an article published in a leading digital health journal [3], two papers presented at major international conferences on medical image analysis [1,2], including one selected for oral presentation [1]; and one paper accepted for oral presentation at an international workshop on interpretability in medical imaging [4]. While the models developed in these studies were primarily evaluated on medical imaging datasets, their underlying architectures can be broadly applicable to a variety of imaging modalities and domains. Full versions of all articles are provided in the Appendix A.

The chapter follows a logical progression. It begins with a sparsity-driven activation constraint to enhance interpretability in disease grading, followed by a clinical evaluation of our inherently sparse interpretable architecture (Sparse BagNet) for efficient and accurate diabetic retinopathy screening. We then present a prototype-based model that combines local and global reasoning. Next, a method is introduced to adapt existing black-box CNN models for self-explainability in high-stakes decisions, and finally, the approach is extended to hybrid architectures that combine the strengths of convolutional neural networks and vision transformers.

The articles in the appendix can be read in the order given below:

1. **Kerol Djoumessi**, Indu Ilanchezian, Laura Kühlewein, Hanna Faber, Christian F. Baumgartner, Bubacarr Bah, Philipp Berens, and Lisa Koch. Sparse activations for interpretable disease grading. In *Medical Imaging with Deep Learning*, 2023. [1]

2. **Kerol Djoumessi**, Ziwei Huang, Laura Kühlewein, Annekatrin Rickmann, Natalia Simon, Lisa Koch, and Philipp Berens. An inherently interpretable AI model improves screening speed and accuracy for early diabetic retinopathy. *PLOS Digital Health*, 2025. [3]

3. **Kerol Djoumessi**, Bubacarr Bah, Laura Kühlewein, Philipp Berens, and Lisa Koch. This actually looks like that: Proto-bagnets for local and global interpretability-by-design. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 718–728, 2024. [2]

4. **Kerol Djoumessi** and Philipp Berens. Soft-CAM: Making black box models self-explainable for high-stakes decisions. *arXiv preprint arXiv:2505.17748*, 2025. [5]

5. **Kerol Djoumessi**, Samuel Mensah, and Philipp Berens. A Hybrid Fully Convolutional CNN-Transformer Model for Inherently Interpretable Disease Detection from Retinal Fundus Images. *Accepted and presented at the International Workshop on Interpretability of Machine Intelligence in Medical Image Computing (IMIMIC)*, 2025. [4]

The first paper introduces *Sparse BagNet* [1], an extension of the original BagNet model that enhances interpretability by producing transparent evidence maps and promoting sparsity in explanations. The model is regularized to focus on the most salient regions, yielding localized and transparent rationales for its predictions. Building on Sparse BagNet, the second study applies it in a retrospective clinical setting to detect early diabetic retinopathy [3]. It evaluates the model's generalizability across multiple publicly available fundus image datasets and assesses the clinical

utility of its explanations in supporting and improving clinician performance. The third work proposes *Proto-BagNet* [2], which combines the BagNet framework with prototype learning [57]. The model provides both local (image-specific) and global (prototypical) explanations, offering interpretable evidence at multiple semantic levels. The fourth article presents *Soft-CAM* [5], a method that transforms standard black-box CNN classifiers into self-explainable models. By incorporating explicit class activation maps into the architecture, the models are inherently interpretable without compromising accuracy. Finally, the fifth paper leverages SoftCAM to introduce a hybrid architecture [4] that combines the spatial inductive biases of CNNs with the long-range dependency modeling capabilities of transformers. This hybrid model is designed for inherent interpretability, providing both high accuracy and explainable outputs suitable for clinical decision support. All results of this thesis are reproducible, with the corresponding code of each project publicly available on GitHub[3].

The remainder of this section is organized as follows: each section corresponds to a specific publication, outlining its motivation, methodology, and key results, followed by a discussion of the main findings and a concise summary of the research contributions. Each section concludes with a brief description of my individual contributions and the venue in which the work was published.

Authors contribution statement

To evaluate my contributions to each paper, my role is summarized using descriptive terms, also presented in tabular form. Specifically, contributions can be categorized as **significant**, **major**, **medium**, or **minor**. To clarify, **significant** denotes a contribution of paramount importance, representing critical foundational work for the project's success. **Major** indicates a crucial involvement without which the submission might not have been possible. **Medium** reflects meaningful contributions that improve the quality of the paper and potentially influence its acceptance. **Minor** refers to supportive input or refinements that improved the paper. This framework provides a transparent and structured guide for communicating the varying degrees of my impact on each published work. The full quantitative contribution details are provided in a separate document alongside the thesis.

3.1 Sparse activations for interpretable disease grading

Published as: Kerol Djoumessi et al. (2023) In *Medical Imaging with Deep Learning (MIDL)*.

3.1.1 Motivation

Interpreting deep learning models often relies on post-hoc saliency maps, which often do not provide actionable feedback or clear insight into the model's decision-making process [43,132,160]. Inherently interpretable models offer a promising alternative for safety-critical applications [33, 34], yet few achieve the high predictive performance of black-box models. The BagNet model [56], an implicit patch-based architecture that independently scores image patches and aggregates them for classification, has demonstrated strong performance on various vision tasks [144, 161]. However, its application in medical imaging—especially ophthalmology—remains limited [145, 146]. Originally, BagNet averages patch features via pooling before classification and generates explanations by scoring each patch individually, then combining these into an attribution map.

[3] All code is publicly available at `https://github.com/kdjoumessi?tab=repositories`

While interpretable, this approach is computationally expensive, often produces cluttered maps with irrelevant activations, and cannot enforce sparsity constraints on the explanations due to its fully connected classifier head. To address these limitations, we proposed *Sparse BagNet*, which preserves BagNet's local feature extraction while removing the average pooling layer and replacing the fully connected classifier with convolutional layers. This architectural change allows the generation of explicit explanation maps directly integrated into the prediction process and enables the use of lasso regularization to promote sparsity, resulting in more sparse and spatially localized explanations.

3.1.2 Results

We validated our approach using the publicly available Kaggle diabetic retinopathy (DR) detection dataset [80], where DR severity ranges from 0 (no DR) to 4 (proliferative DR). Diabetic retinopathy is a microvascular complication related to diabetes and is characterized by progressive retinal lesions that can lead to vision loss [84].

Sparse BagNet was evaluated on two tasks: referable DR detection (binary classification: $0, 1$ vs. $2, 3, 4$) and multiclass DR grading. Despite its inherent interpretability, the model achieved classification performance comparable to a ResNet-50 baseline [65] on both tasks. We then assessed the model's explanations both qualitatively and quantitatively through class evidence maps extracted from the penultimate layer. Compared to GradCAM saliency maps [108] generated post-hoc from the ResNet—which were coarse due to the large receptive field—the sparse BagNet produced significantly sparser heatmaps. Lasso regularization encouraged the model to make decisions based on fewer, more localized retinal regions. Quantitatively, clinical relevance was evaluated using the Top-K precision localization metric on 15 clinically annotated test images. This metric measures the proportion of top-scoring patches overlapping with true lesions, extending the "pointing game" metric [156] to multiple image regions—a critical consideration in DR grading, where disease features are often widespread. Sparse BagNet achieved high Top-K precision (0.791 ± 0.1, SD), significantly outperforming the dense BagNet (0.219 ± 0.1), by focusing almost exclusively on clinically relevant regions. In contrast, the dense BagNet—without sparsity—exhibited lower localization precision, producing more false activations in healthy areas.

3.1.3 Discussion

This work introduced a self-explainable classification model that generates sparse, high-resolution class evidence maps in a single forward pass without compromising predictive performance. Enforcing sparsity reduced false positive activations and simplified model explanations by limiting the contributing regions. The model efficiently produced informative class-wise explanations for multiclass tasks in a single forward pass. Both quantitative and qualitative results demonstrated that Sparse BagNet provides clinically meaningful explanations by accurately localizing DR lesions. For healthy images, the model consistently produced "healthy" heatmaps, characterized by predominantly negative or absent disease evidence, reflecting a low likelihood of disease.

A clinical review of positively activated patches—without prior lesion annotations—revealed that many corresponded to subtle or previously overlooked DR features, highlighting the model's potential to detect early and subtle pathology. Preliminary clinician feedback suggests that Sparse BagNet's interpretable outputs can support validation of predictions, facilitate understanding of failure modes, and build trust in AI-driven decisions. The bounding boxes predicted by the model

were also valuable in highlighting subtle anomalies that might otherwise be missed. These findings underscore the utility of Sparse BagNet as a reliable clinical decision support tool and motivate future human-in-the-loop evaluation of frameworks that allow clinician interaction with the model.

Research contribution: This study introduces Sparse BagNet, an inherently interpretable deep learning model designed to overcome black-box AI limitations in clinical decision-making. Unlike post-hoc methods, it provides human-understandable explanations through explicit class evidence maps. By incorporating a sparsity constraint during training, Sparse BagNet achieved a balance between high predictive performance and transparency. Validated on diabetic retinopathy detection, the model matched state-of-the-art performance while producing localized, clinically meaningful explanations, advancing the development of trustworthy AI in healthcare.

Author contribution: I led this project as first author, co-developing the core idea and methodology with my advisors. I implemented the full codebase, conducted all experiments, analyzed the results, and generated figures. I contributed significantly to the writing, suggested reviewer responses, managed the paper submission, and presented the work at the conference (oral and poster). Additionally, I prepared the preprint and made the final code publicly available[3].

Publication 1: My contributions are summarized as follows:

	Ideas	**Implementation**	**Experiment**	**Analysis**	**Writing**
Kerol Djoumessi	significant	significant	significant	significant	significant

Venue: *Medical Imaging with Deep Learning (MIDL)*, established in 2018, is a leading conference that brings together researchers, clinicians, and industry experts in deep learning and medical imaging. It serves as a platform for advancing automated image analysis in areas such as disease screening, diagnosis, prognosis, treatment planning, and monitoring.

3.2 An inherently interpretable AI model improves screening speed and accuracy for early diabetic retinopathy.

Published as: Kerol Djoumessi et al. (2025) In *PLOS Digital Health*.

3.2.1 Motivation

Diabetic retinopathy screening represents one of the first successful applications of artificial intelligence in medicine [14], providing fast and cost-effective access to care even in settings with limited clinical resources. Several AI systems have received regulatory approval, enabling effective patient triage by distinguishing those requiring specialist attention from those who do not [12, 37], and may also improve screening adherence [14, 15, 84]. However, most state-of-the-art models function as black-boxes [27, 28], offering clinicians limited interpretability and often only binary recommendations [14, 15]. Despite their strong predictive performance, these systems typically require verification by human graders [14, 15], a process that would benefit from transparent explanations of AI decisions. Clinical adoption is further facilitated when clinicians can understand the rationale behind the model recommendations [15, 34, 51]. In prior work, we introduced an inherently interpretable deep learning model—the Sparse BagNet [1]—for DR detection. Preliminary feedback from ophthalmologists indicated that its explanation could help validate predictions, identify

failure modes, and strengthen trust in AI outputs. Building on this foundation and addressing the limited clinical evaluation of self-explainable models and the prevalent lack of transparency in current ophthalmic AI systems, we conducted a formal evaluation of Sparse BagNet for early diabetic retinopathy detection through a retrospective reader study with experienced ophthalmologists.

3.2.2 Results

The Sparse BagNet was trained on the publicly available Kaggle dataset [80] to classify early DR as {grade 0} (no DR) vs. {grades 1–4} (any DR). To assess generalizability, the model was evaluated on ten additional public fundus image datasets spanning diverse cameras, patient populations, and acquisition settings. Five of these datasets included lesion-level annotations, allowing the assessment of explanations through localization performance. In addition, expert lesion annotations were collected for 65 randomly selected fundus images from the Kaggle test set, independently labeled by four ophthalmologists.

Sparse BagNet achieved classification performance comparable to the non-interpretable ResNet-50 baseline [65] in both internal and external datasets, evaluated through accuracy, sensitivity, specificity, and precision. Importantly, it generated class evidence maps that highlighted the discriminative image regions driving its predictions. To assess the clinical relevance of these explanations, bounding boxes were placed around highly activated patches and compared with lesion annotations from both internal and external datasets. Localization performance was benchmarked against two strong post-hoc explanation methods applied to ResNet-50—Guided Backprop [124] and Integrated Gradients [117]—previously shown to perform well on fundus images [162]. Sparse BagNet consistently outperformed these approaches, achieving a localization precision of 0.960 vs. 0.656 (Guided Backprop) on the internal dataset, and 0.965 vs. 0.249 on the external DDR dataset [163]. Analysis of false positives (images misclassified as DR with high confidence > 0.75) revealed that most activated patches contained subtle anomalies related to DR, such as microaneurysms and exudates that were below diagnostic thresholds. The review by two clinicians confirmed these findings, suggesting that the model may detect early pathological signals overlooked by manual annotations, thereby also highlighting potential incompleteness in ground truth labels.

Finally, a retrospective reader study with six experienced ophthalmologists evaluated the clinical impact of AI support under three conditions: absence of AI assistance ("H"), AI prediction with confidence only ("H+AI"), and AI prediction with localized explanations ("H+XAI"). The average diagnostic accuracy improved from 0.611 (H) to 0.758 (H+AI) and further to 0.786 (H+XAI). The gains were more pronounced in mild DR cases (grade 1), where accuracy increased from 0.483 (H) to 0.617 (H+AI) and 0.733 (H+XAI). For healthy cases, any AI support substantially improved performance (H: 0.567; H+AI: 0.842; H+XAI: 0.817). Beyond accuracy, explanations accelerated decision-making: average times were $15.2s$ (H), $15.9s$ (H+AI), and $11.7s$ (H+XAI), with the largest speed-up again observed in mild DR cases (H: $15.2s$; H+AI: $17.5s$; H+XAI: $12.1s$). Together, these results demonstrate that interpretable AI support from Sparse BagNet improves both accuracy and efficiency, particularly in challenging early DR detection—grade 1, a challenging task even for experienced ophthalmologists.

3.2.3 Discussion

This study demonstrated that the inherently interpretable Sparse BagNet achieves classification performance comparable to a black-box baseline model on internal and ten external datasets, showing strong zero-shot generalization across diverse devices, populations, ethnicities, and imaging conditions. Unlike conventional DR screening models that provide only binary outputs, Sparse BagNet generates highly precise evidence maps that localize image regions driving its predictions. Interestingly, despite being trained solely with image-level labels (without pixel-level annotations), the highlighted regions consistently aligned with clinically relevant DR lesions such as microaneurysms, drusen, or hemorrhages. Even when predictions diverged from reference labels, explanations often revealed clinically suspicious features, providing additional diagnostic insight.

The retrospective reader study further confirmed that providing ophthalmologists with visual explanations significantly improved grading accuracy, particularly for challenging cases, while also reducing decision time. These results demonstrate that interpretable AI support can improve screening performance. The effectiveness of Sparse BagNet's explanations suggests that embedding such models into AI-driven workflows could strengthen human verification and foster greater clinician trust in automated recommendations [15, 74]. Unlike prior human-AI studies reporting adverse effects of model errors on clinician decisions [101], our findings did not show any such negative influence. Future work could include prospective clinical trials to assess the impact of explainable AI on screening quality and efficiency in real-world practice, with potential extensions to other diagnostic tasks, such as breast cancer and pneumonia detection [17, 24].

Research contribution: This work demonstrated that deep learning models can achieve inherent interpretability without compromising predictive performance, addressing a major barrier to clinical adoption of AI systems. Leveraging Sparse BagNet architecture, the model produced transparent and reliable explanations that improved diagnostic accuracy by up to 17.5% for mild diabetic retinopathy and reduced the screening time by approximately $\approx 25\%$. These inherently interpretable outputs can be seamlessly integrated into existing clinical workflows and reporting systems, fostering transparency, clinical trust, and earlier, more accurate diagnoses in applications such as diabetic retinopathy screening.

Author contribution: I led this project as the first author, co-developed the study in collaboration with my PhD advisors. Together, we formulated the research question and designed the clinical user study. I implemented the full codebase, conducted all experiments, including the user study analysis, and generated most of the figures. I was heavily involved in manuscript writing and journal formatting, and also provided feedback on the study's software[4]. In addition, I prepared the preprint version, uploaded the final code to GitHub[3] to ensure public accessibility, and coordinated the overall project execution.

Publication 2: My contributions are summarized as follows:

	Ideas	**Implementation**	**Experiment**	**Analysis**	**Writing**
Kerol Djoumessi	significant	significant	significant	significant	significant

Venue: *PLOS Digital Health* is an open-access, interdisciplinary journal that publishes innovative research that uses digital tools, technologies, and data science to advance human health.

[4] Available at `https://github.com/berenslab/retimgtools/releases/tag/v1.1.0`

Committed to open science, the journal promotes equitable and unbiased healthcare through ethically conducted and impactful studies. It brings together contributions from a global community of engineers, clinicians, researchers, social scientists, and industry leaders.

3.3 This actually looks like that: Proto-BagNets for local and global interpretability-by-design

Published as: Kerol Djoumessi et al. (2024) In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*.

3.3.1 Motivation

Post-hoc explainability methods often fail to reflect a model's actual reasoning process. Inherently interpretable approaches—such as BagNets [1,56], concept-based models [107,118], and prototype-based networks [57,134]—offer a promising alternative. BagNets leverage small receptive fields to provide fine-grained explanations [1], but lack global interpretability [147]. Concept-based models [107] incorporate human-aligned predictions at a higher level, enabling globally interpretable predictions, but typically depend on predefined concept sets [118, 127]. Prototype-based networks [57], a subclass of concept-based models, overcome this by learning prototypes directly from training data, and classifying inputs based on similarity to them, yielding both local (via similarity maps) and global (via prototype visualization) explanations. Although prototype-based models are gaining attention [134, 136, 141] for their intuitive, human-aligned reasoning, their use in medical imaging remains limited [16, 73, 134]. A key limitation lies in their receptive fields, which produce overly broad prototypes [141,142], reducing their ability to capture subtle but clinically relevant features, particularly when they are relatively small compared to the image size. To address this gap, we introduced Proto-BagNet, which combines BagNet's localized interpretability with the global reasoning capabilities of prototype learning, delivering interpretable and faithful explanations at both local and global levels.

3.3.2 Results

Proto-BagNet was validated on 2D retinal OCT images for drusen detection using a publicly available dataset [81]. Drusen are subretinal deposits and key indicators of retinal diseases such as age-related macular degeneration (AMD) [81,85]. In OCT scans, their size, number, and location are critical for disease staging and guiding early interventions to prevent vision loss [85].

Proto-BagNet integrates several interpretability and performance improvements, including a soft aggregation module to prevent interference from cross-class prototype [158], a top-K similarity scoring mechanism to capture multiple relevant regions [73], a sparsity constraint to focus on disease-relevant features [1,120], and a novel dissimilarity loss to promote diverse non-redundant prototypes. Classification performance was benchmarked against ProtoPNet (ResNet-50 backbone) [57] and non-prototype models (dense BagNet [1] and ResNet-50 [65]). Proto-BagNet achieved comparable classification performance, demonstrating that its interpretability enhancements did not compromise performance. For interpretability evaluation, ProtoPNet's large prototypes (covering most of the retina, receptive field 427×427) were contrasted with Proto-BagNet's fine-grained prototypes (33×33), which localized clinically relevant drusen features. An independent ophthalmologist review confirmed their clinical significance. A small set of 40 annotated

test images for drusen was used to assess the localization precision of prototypical activations, showing that the disease prototypes activated on these images effectively captured and localized drusen-related lesions with high precision (0.84 ± 0.20), demonstrating strong alignment with clinically meaningful features. Faithfulness was assessed by masking all but the top-5 prototypical regions, which nearly preserved classification performance (AUC: 0.9918 vs. 0.9916), confirming that predictions relied exclusively on these regions and that the explanations accurately reflected the model's internal decision process.

3.3.3 Discussion

We introduced Proto-BagNet, a novel prototype-based interpretable-by-design model that provides highly localized, faithful explanations alongside global interpretability through meaningful prototypes. Clinical evaluation by an ophthalmologist confirmed that the learned prototypes captured diverse and clinically relevant features and accurately localized drusen lesions in OCT images. Interpretability-enhancing components—soft-aggregation, top-K selection, sparsity, and dissimilarity loss—may slightly trade off classification performance, emphasizing the importance of task-specific parameter tuning. This highlights the importance of tuning some parameters according to the clinical context. Unlike standard classification tasks with large, centered objects, medical imaging often involves small, spatially distributed features [17,85,163], requiring attention to multiple regions. For instance, drusen and some diabetic retinopathy lesions are relatively small and scattered [85,163], requiring prototype-based models to consider multiple regions rather than focusing on the most relevant area. The optimal receptive field is also task-dependent: in this study, 33×33 patches effectively captured drusen features [85], but the size of the receptive field can be adapted based on clinical knowledge or image resolution. Regularizing the similarity maps enforced focus on relevant regions, while the dissimilarity loss reduced redundancy, promoting diversity, and improving interpretability. Nevertheless, the lack of global attention in architecture with limited receptive fields can constrain predictive performance in tasks where important features span larger areas [147], such as advanced DR and pneumonia detection [1,17]. Furthermore, prototype-based approaches can be challenging to train effectively, highlighting the need for more easily trained, generalizable, and model-agnostic architectures.

Research contribution: This work introduces Proto-BagNet, a prototype-based model specifically designed for fine-grained detection in medical imaging that addresses key limitations of prior approaches, including imprecise explanations and redundant prototypes [140–142]. By leveraging BagNet's small receptive field, Proto-BagNet learns informative, fine-grained prototypes that provide precise and faithful explanations without sacrificing predictive performance, thereby advancing interpretability and trust in clinical AI. The model incorporated constraints to promote prototype diversity and achieved high explanation faithfulness, representing a significant advance toward more interpretable and trustworthy clinical AI systems.

Author contribution: I led this project as the first author, co-developing the core idea with my PhD advisors. I implemented the codebase, performed all experiments, analyzed the results, and generated figures. I contributed substantially to the manuscript and drafted responses to the reviewer's comments. Additionally, I managed the paper submission, presented the work at the conference (poster), prepared the preprint, and made the final code publicly available on GitHub[3].

Publication 3: My contributions are summarized as follows:

	Ideas	**Implementation**	**Experiment**	**Analysis**	**Writing**
Kerol Djoumessi	significant	significant	significant	significant	significant

Venue: The *Medical Image Computing and Computer-Assisted Intervention (MICCAI)* conference is the leading international forum for medical image computing, machine learning in medical imaging, and computer-assisted interventions and robotics. Established in 1998, it fosters the exchange of cutting-edge knowledge, expertise, and experiences among leading scientists, clinicians, and educators from top academic, clinical, governmental, and industry institutions worldwide.

3.4 Soft-CAM: Making black box models self-explainable for high-stakes decisions

Published as: Kerol Djoumessi and Philipp Berens (2025), *arXiv preprint.*

3.4.1 Motivation

Convolutional neural networks (CNNs) are widely used in high-stakes domains such as medicine, often surpassing human performance [6]. Yet, their limited interpretability poses a major barrier to adoption, where transparency and trust are critical. Post-hoc methods such as class activation maps (CAM) [111] and their numerous variants [108,112,117,124] attempt to approximate model reasoning after training, but their explanations are often sensitive, unreliable, and not very faithful to the underlying decision process. To address these shortcomings, inherently interpretable models have been proposed [34], embedding interpretability directly into the architecture [57, 107, 149], yielding more faithful and trustworthy explanations [106]. However, these models typically require specialized designs [56, 57], limiting their generalization and compatibility with standard CNNs. Motivated by this gap, we proposed SoftCAM, a simple yet effective method that makes back-box CNN architectures inherently interpretable. By removing the global average pooling (GAP) layer and replacing the fully connected head with a convolution-based class evidence layer, SoftCAM preserves spatial information and generates explicit class activation maps that directly inform predictions. This design also supports ElasticNet penalties to tailor explanations to specific tasks.

3.4.2 Results

We evaluated SoftCAM on three public medical imaging datasets covering different modalities: Kaggle Diabetic Retinopathy [80], Retinal OCT [81], and RSNA Chest X-Ray (CXR) [79]. Each dataset included lesion annotations for subsets of images, allowing an objective evaluation of the explanations. SoftCAM was implemented on two standard CNN backbones—ResNet-50 [65] and VGG-16 [63]—which differ in their classification head: ResNet uses a single fully connected layer, whereas VGG uses multiple layers. For comparison, we benchmarked SoftCAM against popular post-hoc saliency approaches: gradient-based methods (GradCAM [108], LayerCAM [109]), a gradient-free method (ScoreCAM [112]), and backpropagation-based techniques (Integrated Gradients [117], Guided Backpropagation [124]), each on their respective black-box models.

To introduce inherent interpretability, we systematically replaced the fully connected heads of ResNet and VGG models with convolutional layers and applied lasso regularization to the class-evidence layers. Both sparse (lasso) and dense (no penalty) SoftCAM variants preserved

classification performance comparable to the original backbones. Qualitative comparisons revealed that SoftCAM—particularly the sparse variant—produced sharper and more interpretable evidence maps, with high-activation regions aligning closely with annotated lesions. Quantitative assessment confirmed these findings: SoftCAM consistently outperformed post-hoc methods in both lesion localization precision and faithfulness to model decision-making. Given the larger bounding boxes in CXR compared to retinal datasets, we extended the evaluation by introducing *activation sensitivity*, which measures the proportion of explanation within ground-truth bounding boxes. This complements the activation precision [73] by focusing on the model to avoid false negatives. Across datasets, SoftCAM achieved superior performance in both precision and sensitivity. Finally, we investigate the impact of different regularization strategies. Lasso promoted sparsity by eliminating irrelevant activations, producing cleaner maps, while ridge retained small values, yielding denser but less focused explanations. Despite these differences, all SoftCAM variants outperformed post-hoc methods across metrics. In multi-class tasks, SoftCAM generated class-specific explanations in a single forward pass, achieving state-of-the-art predictive performance while improving efficiency and explanation faithfulness compared to post-hoc saliency techniques.

3.4.3 Discussion

SoftCAM provides a straightforward yet effective protocol for transforming standard black-box CNNs into inherently interpretable models. Evaluation in various medical imaging tasks—including retinal and chest X-rays images—using two widely adopted backbones, ResNet-50 [65] and VGG-16 [63], showed that SoftCAM preserves classification performance while delivering more faithful and precise explanations than post-hoc methods. By replacing the classification head with a convolution-based class evidence layer, SoftCAM generates explicit class activation maps in a single forward pass, enabling efficient and trustworthy explanations. Its design supports ElasticNet regularization, allowing users to balance localization precision and sensitivity without significantly sacrificing predictive accuracy, and in some cases even improving it. The results confirmed that convolutional classifiers can simultaneously achieve high predictive performance and interpretability, making SoftCAM a compelling solution for self-explainable CNN models. Beyond predictive performance, it provides valuable insights into model decision-making, helping to identify errors and spurious correlations in standard CNNs without relying on post-hoc attribution methods. Future work may extend SoftCAM to additional datasets and CNN architectures, as well as to other black-box models, such as hybrid CNN-Transformer designs [31, 32].

Research contribution: This work introduced SoftCAM, a simple yet effective approach that transforms standard CNNs into inherently interpretable models by replacing their classification head with convolutional layers. Extensive experiments on medical imaging tasks showed that SoftCAM preserves high predictive performance while producing faithful, class-specific explanations. Thanks to ElasticNet regularization, which enables task-specific control over precision and sensitive localization, offering an efficient and trustworthy self-explainable alternative to black-box CNNs.

Author contribution: I led this project as the first author and came up with the core idea. I implemented the full codebase, conducted all experiments, analyzed the results, and generated the figures. I contributed substantially to writing the manuscript and managed submission to arXiv, in addition to uploading the final code to GitHub[3] to ensure reproducibility and public access.

Publication 4: My contributions are summarized as follows:

	Ideas	**Implementation**	**Experiment**	**Analysis**	**Writing**
Kerol Djoumessi	significant	significant	significant	significant	significant

Venue: *ArXiv* is the world's largest open-access repository for scientific research, widely used in fields such as mathematics, and computer science, and beyond. It enables rapid dissemination of preprints and working papers without the delays of traditional peer review, fostering open collaboration, accelerating scientific discovery, and serving as an essential resource for researchers.

3.5 A Hybrid fully convolutional CNN-Transformer model for inherently interpretable disease detection from retinal fundus images

Published as: Kerol Djoumessi et al. (2025) In *International Workshop on Interpretability of Machine Intelligence in Medical Image Computing (IMIMIC) at MICCAI.*

3.5.1 Motivation

Convolutional neural networks (CNNs) are widely used in medical imaging for their ability to hierarchically extract local features. While effective in capturing fine spatial features, CNNs are often constrained by limited receptive fields [56, 147], making them less effective at modeling long-range dependencies between image regions. Vision transformers (ViTs) [64] overcome this limitation through self-attention (SA) mechanisms [70], enabling global context modeling. However, ViTs lack the inherent spatial localization provided by convolutions, demand large-scale datasets and substantial computational resources [31, 32], and introduce new challenges for interpretability [30–32]. Hybrid CNN–ViT architectures aim to combine the spatial precision of CNNs with the global attention of transformers [30, 31]. Despite promising results, their interpretability remains limited [30, 32, 164], often relying on post-hoc tools, including transformer-specific visualization techniques. Existing interpretability approaches extend CNN-based attribution techniques to ViTs [117, 125] or design ViT-specific explanation methods [165, 166]. Both are typically post-hoc and may not faithfully capture the model's decision process. To address these challenges, we proposed an interpretable hybrid fully convolutional CNN-Transformer architecture for medical image analysis, with a focus on the detection of retinal disease. Unlike standard post-hoc saliency methods for ViT-based models, our approach generates faithful, localized class-evidence maps directly integrated into the prediction pipeline, eliminating the need for external explanation techniques.

3.5.2 Results

We validated our approach on two clinically relevant tasks: Diabetic Retinopathy (DR) detection and Age-Related Macular Degeneration (AMD) severity classification, using publicly available color fundus image datasets [80, 167]. The proposed hybrid model integrates recent advances to enhance both performance and interpretability. It combines dual-resolution self-attention (DRSA) [164] to model multi-scale (fine- and coarse-grained) dependencies with convolutional vision transformers (CvTs) [168], replacing linear attention layers with convolutions to preserve spatial information. Built-in interpretability was achieved by replacing the standard classifica-

tion head with a convolutional class-evidence layer [1, 5], allowing sparsity regularization for more localized and clinically relevant explanations.

We evaluated the framework using ResNet-50 [65] and BagNet-33 backbones [56], both with and without sparsity, and compared it against state-of-the-art CNNs (ResNet-50 [65], dense BagNet [1]) and transformer-based models (ViT [64], Swin [71]). Ablation studies replacing dual-resolution SA with fully connected layers (FCL) confirmed its critical contribution. Despite sparsity constraints, our interpretable hybrid models achieved state-of-the-art predictive performance on both tasks. Qualitative comparisons with GradCAM [117] on the ViT baseline showed that it produced cluttered, hard-to-interpret heatmaps, while our class-evidence maps clearly highlighted class-relevant retinal features. Sparsity further refined the focus on smaller, clinically meaningful regions. Quantitatively, we assessed the lesions' localization capability of our models on the IDRiD dataset [143] by measuring the proportion of positively activated regions that overlap with annotated lesions [1]. The sparse BagNet–Transformer achieved the highest precision, outperforming all baselines—including the dense BagNet and GradCAM applied to ViT—indicating that convolutional attention improved both classification and interoperability. Faithfulness was assessed by removing top-ranked patches from evidence maps and measuring the drop in class confidence [155]. For DR detection, the sparse BagNet–Transformer showed the highest faithfulness, while for AMD severity classification, the sparse ResNet-Transformer performed best. This difference likely reflects CNNs' larger receptive fields, which better capture the broader lesion patterns typical of AMD, highlighting differences in lesion sizes between tasks. Finally, class-specific evidence maps visualizations on the Kaggle DR dataset demonstrated that the sparse BagNet–Transformer produced localized explanations, closely aligned with prediction and relevant features, while dense models generated broader, less informative explanations.

3.5.3 Discussion

We presented the first inherently interpretable hybrid CNN-Transformer architecture for medical image classification, demonstrated on DR detection and AMD severity classification using retinal fundus images. The model is backbone-agnostic, allowing the selection of architectures suited to specific disease characteristics. Two CNN backbones were evaluated: ResNet-50 [65], which capture global spatial patterns relevant to AMD, and BagNet [56], emphasizing fine-grained local features critical for DR. The hybrid ResNet–Transformer yielded coarser, yet informative evidence maps aligned with larger AMD lesions, while the BagNet–Transformer produced more localized explanations suitable for DR detection. The DRSA effectively expanded BagNet's limited receptive field and enhanced ResNet's ability to model long-range multi-scale dependencies. Unlike typical hybrid models with fully connected classifier heads, our architecture used an explicit class-evidence layer that outputs spatial heatmaps directly tied to predictions [1, 5], allowing interpretation without post-hoc methods. All variants achieved comparable predictive performance, with the sparse BagNet–Transformer providing the most informative explanations for DR detection, while the sparse ResNet–Transformer performed best for AMD. The model produced class evidence maps and predictions in a single forward pass, in contrast to post-hoc methods like GradCAM, which require an additional backward pass for each class. Strong alignment was observed between attention and evidence maps, particularly in sparse variants, highlighting the model's ability to interpret long-range dependencies. Consistent with previous findings [1], increased sparsity occasionally reduced sensitivity to late-stage DR, likely due to underrepresentation in training

data. Although validated on fundus images, this framework can generalize to other medical imaging tasks and modalities, with the potential to incorporate diverse CNN backbones and attention mechanisms.

Research contribution: This work presents, to the best of our knowledge, the first inherently self-explainable, fully convolutional hybrid CNN-Transformer for image analysis. The proposed approach preserves spatial information and provides built-in interpretability while effectively capturing long-range dependencies across multiple resolutions. Unlike prior hybrid models that rely on post-hoc explanations [78, 169] or use complex self-explainable frameworks [170], our method offers a simpler and more effective solution for clinical imaging applications.

Author contribution: I led this project as the first author, developing the core idea and implementing the codebase, conducting experiments, analyzing results, and generating figures. I contributed extensively to manuscript writing and prepared responses to reviewer comments. Additionally, I managed the paper submission, presented the work at the Workshop (both poster and oral sessions), prepared the preprint, and uploaded the final code to GitHub[3] for public access.

Publication 5: My contributions are summarized as follows:

	Ideas	**Implementation**	**Experiment**	**Analysis**	**Writing**
Kerol Djoumessi	significant	significant	significant	significant	significant

Venue: *The Interpretability of Machine Intelligence in Medical Image Computing (IMIMIC) workshop*, held annually since 2018 as part of the MICCAI conference series with proceedings, focuses on advancing the transparency and interpretability of machine learning models in medical imaging. It brings together researchers and clinicians to explore methods that improve model understanding, foster trust, and support the safe clinical deployment of AI in healthcare.

4 Discussion and conclusion

In this thesis, we addressed the challenge of opacity in supervised deep learning classifiers for image analysis, with a particular emphasis on medical imaging. This final chapter first summarizes the findings, highlights the strengths and limitations of the proposed models, and then outlines directions for future research. More specifically, it explores how the presented approaches tackle key interpretability challenges, offering practical value for both general computer vision tasks and high-stakes applications such as clinical decision support.

4.1 Summary and contributions

Deep learning has achieved remarkable success in medical image analysis, often matching or even surpassing expert-level [12, 21, 37], driving a growing interest in AI-assisted decision-making for high-stakes clinical applications [12, 37, 49]. However, the widespread adoption of such systems remains limited by their lack of transparency [33, 34, 51]. Most deep learning models operate as black boxes, offering little insight into their reasoning, which poses barriers to clinical trust, interpretability, and regulatory approval [38, 44]. Post-hoc explanation methods attempt to mitigate this opacity by approximating models' reasoning [27, 28, 126], using various approaches including gradient-based [108, 109], gradient-free [111, 112], perturbation-based [114, 115], concept-based [118, 119], and example-based techniques [28, 121]. Although these methods are widely used in and beyond medical imaging, they remain limited by key shortcomings [43, 132, 160]: their explanations can be unstable, unfaithful to the model's true decision process, and sensitive to spurious correlations.

Inherently interpretable models provide a compelling alternative for safety-critical domains [34, 106], as they integrate interpretability directly into the model architecture, making explanations an intrinsic feature rather than an optional post-hoc add-on. This thesis contributes to this approach by developing self-explainable classifiers and adapting existing black-box architectures to incorporate built-in interpretability. The proposed models include:

- **Sparse BagNet** [1], which enforces sparsity in its explanations, encouraging the model to focus on localized, clinically meaningful image regions;
- **Proto-BagNet** [2], integrating BagNet's localized receptive fields with prototype learning to provide local and global interpretability;
- **SoftCAM** [5], a lightweight protocol that modifies standard CNNs to produce interpretable, class-specific evidence maps in a single forward pass; and
- **Hybrid CNN-Transformer** [4], the first fully convolutional and inherently interpretable CNN-Transformer hybrid for medical imaging.

Some proposed approaches, such as Sparse BagNet and Proto-BagNet, extend existing interpretable baselines (e.g., BagNet [56], ProtoPNet [57]), while others combine complementary strengths (e.g., Proto-BagNet, Hybrid CNN-Transformer) or introduce effective modifications to standard CNNs to make them inherently explainable and more transparent (SoftCAM, Sparse BagNet). Together, these contributions provide a foundation for developing more trustworthy and interpretable AI systems for image analysis and promote a rigorous, meaningful evaluation of model explainability.

4.1.1 Sparse BagNet and applications

Built on the inherently interpretable BagNet architecture [56], Sparse BagNet [1] improves transparency by generating explicit class evidence maps that explain its predictions. By combining deep feature extraction with the interpretability of simpler linear models, a sparsity constraint is applied in the loss function, promoting localized and clinically meaningful explanations without sacrificing performance. The model was initially validated on diabetic retinopathy detection for both multiclass classification and referable DR screening, achieving results comparable to black-box baselines while providing interpretable outputs. Collaboration with ophthalmologists showed that Sparse BagNet could identify early and subtle signs of presymptomatic DR—often challenging even for experts—, leading to an 17.5% increase in diagnostic accuracy and roughly a 25% reduction in the screening time for difficult cases (grade 1, corresponding to midl detection) [3].

Sparse BagNet was later extended to other ophthalmology applications, including Epiretinal Membrane (ERM) detection from OCT images [59] and age-related macular degeneration (AMD) progression prediction from color fundus images [58]. For ERM detection, the model achieved performance comparable to a ResNet-based black-box baseline, generalized effectively to external datasets, and, in a multitask setting, identified additional related pathologies. A user study confirmed that its visual explanations were consistent with clinical reasoning. In AMD progression modeling, Sparse BagNet was used to parameterize the hazard function [171] for deep survival analysis, accurately predicting the risk of developing late-stage AMD. Salient image regions highlighted by its sparse evidence maps closely aligned with established clinical markers and outperformed standard post-hoc saliency methods applied to black-box models.

Finally, the dense BagNet model was submitted to the MIDRC XAI Challenge 2024[5], which focused on classifying pneumonia-related opacities in chest radiographs. Despite being self-explainable and trained with minimal supervision, it achieved the 7th place[6], outperforming most approaches based on post-hoc explanations. Notably, many top-performing solutions relied heavily on ground-truth segmentations and lacked inherent transparency, highlighting dense BagNet's ability to achieve competitive performance while providing interpretable outputs without additional annotation supervision.

However, while dense and sparse BagNet improve transparency through local attribution maps, they remain limited in capturing global disease patterns. Furthermore, their constrained receptive field can reduce performance in tasks where relevant lesions exceed the patch size, such as pneumonia detection or advanced diabetic retinopathy.

4.1.2 Proto-BagNet: a step forward in prototype learning

Prototype-based models have emerged as a promising framework for self-explainable systems [57, 134, 136], yet their application in medical imaging—particularly in retinal diagnosis—remains limited compared to applications in natural image analysis [136, 144, 158]. Earlier models faced several challenges, most notably large receptive fields [141, 142], which led to low-resolution activation maps due to excessive downsampling and feature aggregation in the backbone network. These coarse similarity maps, typically upsampled to match the input resolution, often produce imprecise spatial explanations [141, 142], commonly visualized by drawing bounding boxes around the most discriminative image region [57]. The combination of large receptive fields and naive

[5]MIDRC XAI Challenge: `https://qtim-challenges.southcentralus.cloudapp.azure.com/competitions/36/`
[6]MIDRC XAI Challenge rankings: `https://www.midrc.org/xai-challenge-2024`

upsampling frequently obscured subtle localized disease features, highlighting the critical importance of receptive field size in prototype-based interpretability [136, 141]. Additionally, previous models often exhibited redundant and semantically irrelevant prototypes [136, 140], thus reducing both interpretability and clinical relevance.

Proto-BagNet addressed these challenges by leveraging BagNet's inherently small receptive field to generate high-resolution activation maps, allowing precise and localized explanations without the need for post-hoc methods. It further enhances interpretability through a dissimilarity constraint that promotes diverse, non-redundant, and semantically meaningful prototypes. Consequently, its learned prototypes are both disentangled and clinically representative, offering faithful and transparent insight into the model's reasoning. Faithfulness analysis confirmed that Proto-BagNet's explanations accurately reflect its internal decision process, representing a significant advancement towards trustworthy, interpretable prototype-based models in medical imaging—where relevant clinical signals are often subtle and spatially localized.

Compared to Sparse BagNet, which provides only local explanations for individual predictions, Proto-BagNet extends interpretability by offering both global and local explanations. Globally, it replaces the learned prototypes from the embedding space with their closest patch representations in the training set and retrains the classification head using the features of these prototypes, enabling a global understanding of class-level reasoning. Locally, it computes patch-wise similarity between each prototype and the test image, producing fine-grained attribution maps—termed prototype similarity maps—that highlight relevant image regions.

However, despite offering complementary local and global interpretability, both Proto-BagNet and Sparse BagNet shared an inherent limitation: their constrained receptive fields limit the ability to capture large lesions, which can negatively impact classification performance. Furthermore, prototype-based models present additional training challenges: the prototype replacement step requires retraining the classification layer, increasing training complexity and computational cost. Their architecture design also often demands task-specific optimization strategies and careful hyperparameter tuning, such as the number of prototypes per class and the receptive field size, which can affect model convergence and hinder scalability across datasets and clinical applications.

4.1.3 Fully convolutional architectures for advancing model explainability

Convolutional neural networks are widely used in high-stakes vision tasks due to their strong inductive biases—such as locality and translation invariance—which make them effective at capturing fine-grained image features [21, 66]. However, standard CNN architectures typically rely on fully connected layers in their classifier heads, which obscure the decision-making process and require post-hoc interpretability methods. This opacity limits their applicability in safety-critical domains where transparent and human-understandable reasoning is essential. With regulatory frameworks such as the European AI Act and the General Data Protection Regulation increasingly mandating explainability [38, 44, 47], models must not only achieve high accuracy, but also provide clear and reliable insights into their predictions. Beyond compliance, explainability helps uncover model shortcuts, highlight relevant features, and foster trust in real-world applications [12, 50, 51].

Vision Transformers [30, 64, 71, 168] have emerged as strong alternatives to CNNs in both general and medical imaging tasks [30, 32, 67]. However, they face similar interpretability challenges, as their transformer blocks and classification heads often rely on linear multilayer perceptrons that do not preserve spatial locality. Even convolutional variants, such as Convolutional Vision

Transformers [168] and Swin Transformers [71], incorporate inductive biases for improved performance, but still rely on post-hoc explanations due to their inherently opaque architectures, thereby limiting their clinical applicability.

To address these issues, we showed that both CNN and hybrid CNN-Transformer models can achieve inherent interpretability by leveraging the theoretical equivalence between the linear and 1×1 convolutional layers [172]. Fully connected layers can be reformulated as 1×1 convolutions, preserving model complexity while producing spatially resolved class evidence maps [1, 172]. By replacing global average pooling and fully connected classifier layers with convolutional operations, standard CNNs can be converted to fully convolutional networks that preserve spatial information, eliminating the need for approximate post-hoc explanations [1, 5]. This principle was further extended to hybrid CNN-Transformer models [4] by substituting linear layers in attention blocks and classifier heads with convolutional counterparts, maintaining performance while producing inherently interpretable outputs. When using BagNet [56] as the CNN backbone, the resulting model generates high-resolution evidence maps that provide fine-grained, spatially precise explanations, which can be regularized during training to further enhance interpretability [5].

Overall, this framework introduces a practical and generalizable protocol for building inherently interpretable vision models—spanning CNNs and hybrid CNN-Transformer—designed for safety-critical domains. By removing reliance on post-hoc methods while maintaining strong predictive performance, it advances the development of trustworthy AI systems capable of transparent decision-making. Moreover, integrating CNNs with transformers mitigates BagNet's localized receptive field limitation by enabling the model to capture long-range dependencies through self-attention. However, the granularity of the resulting attribution maps remains dependent on the receptive field size: models with smaller receptive fields, such as BagNet, produce fine-grained explanations, whereas those with larger receptive fields, such as ResNet, yield coarser attributions even under sparsity constraints.

4.2 Outlook and future work

In this thesis, we introduced inherently self-explainable deep learning classifiers for medical image analysis, namely Sparse BagNet [1], Proto-BagNet [2], SoftCAM [5], and a hybrid CNN-Transformer model [3]. Although primarily evaluated in medical imaging, the underlying principles and architecture can be generalized to broader vision tasks. Through this work, we demonstrated that (1) the interpretability of existing self-explainable models can be systematically enhanced for greater transparency and usability; (2) combining complementary strengths from different self-explainable architectures within a unified framework further improves interpretability; and (3) standard black-box CNNs can be systematically transformed into inherently interpretable models without sacrificing predictive performance.

Despite these advances, several opportunities remain for further refinement and expansion. Future research could focus on improving generalizability across heterogeneous datasets, scaling to complex clinical workflows, and integrating multimodal data or explicit domain knowledge. Advancing in these directions will be essential to enhance both accuracy and interpretability, ultimately fostering the deployment of self-explainable AI systems in real-world healthcare settings.

Clinical validation and broader applications of Sparse BagNet in medical imaging. The sparse BagNet model has demonstrated strong potential to support clinicians in medical

imaging, as shown by its successful retrospective application in multiple ophthalmic tasks, including diabetic retinopathy detection [1, 3], age-related macular degeneration progression modeling [58], and epiretinal membrane detection from OCT images [59]. While these studies involved clinicians in the evaluation process, they were conducted retrospectively, underscoring the need for prospective investigations to rigorously evaluate generalization, clinical utility, and impact in real-world—particularly for diabetic retinopathy and other retinal diseases. Beyond ophthalmology, Sparse BagNet's architectural design makes it well-suited for medical imaging tasks featuring small, localized disease patterns that align with its receptive field. Promising areas of application include skin lesion classification in dermoscopy [101, 145, 169]; disease detection in thoracic imaging [16, 17, 72, 138]—such as lung nodule or pneumonia detection in chest X-ray and CT scans—neurological disorders identification from brain imaging [78, 139, 147], and breast cancer detection from mammography and related modalities [24, 73]. Comprehensive retrospective and prospective validation across these diverse domains will be essential not only to establish Sparse BagNet as a generalizable and interpretable diagnostic model but also to uncover domain-specific limitations and guide future architectural and methodological refinements.

Expanding and enhancing explanations in Sparse BagNet. A key limitation of the Sparse BagNet lies in its reliance on visual evidence maps as the primary form of explanation. Although these effectively localize discriminative regions, their interpretation often requires clinical expertise and may not be intuitive for non-expert users. Several complementary strategies could address this limitation. One promising direction involves integrating Automated Concept-based Explanation (ACE) [119], which can identify semantically meaningful concepts from the patch-wise features space of a pretrained Sparse BagNet. The resulting concept representative cluster can then be linked to class predictions using techniques such as TCAV [120] to quantify the relevance of each concept, thereby producing global post-hoc summaries that complement the model's local attribution maps. Another approach involves annotating and classifying the high-attribution patches using an auxiliary CNN trained on lesion-level annotations—available in several diabetic retinopathy datasets [143,163]—to improve the clinical clarity and granularity of the explanations. Beyond purely visual approaches, large language models (LLMs) offer a compelling avenue for translating evidence maps and predictions into coherent textual explanations, as demonstrated in recent works such as GraphXAIN [98] and LLMExplainer [99]. In such frameworks, an LLM could serve as a decoder, taking as input the image, evidence map, predicted, and true labels to generate natural-language justifications or clarifications of misclassifications. Finally, interpretability could be further enriched through diffusion-based counterfactual generation [90], wherein targeted modifications to high-attribution regions simulate alternative outcomes—enabling contrastive reasoning "what-if" and offering a more complete picture of model behavior.

Prototype learning: challenges and perspectives. Prototype learning has emerged as another promising paradigm for self-explainable models [134, 136], but its systematic application and evaluation in medical imaging remain limited [134], leaving unresolved challenges and questions about clinical utility. A primary limitation lies in the model complexity, as numerous hyperparameters—such as the number of prototypes, receptive field size, and component-specific training settings—must be carefully tuned. This complicates optimization, hinders transferability to new tasks, and increases the risk of suboptimal performance. While human-centered evaluations have been explored in natural image domains [173], clinical validation involving healthcare profes-

sionals remains difficult, limiting practical applicability. Among these challenges, receptive field size is particularly influential in determining the spatial precision and faithfulness of prototype explanations. This parameter depends on both the choice of backbone architecture and the chosen prototype-based variants (e.g., ProtoPNet [57], ProtoCAPs [138], ProtoEval [158]), thereby requiring multi-level tuning. Such intertwined dependencies make prototype-based models difficult to adapt across tasks, as adjusting one hyperparameter can cascade into changes in both performance and interpretability. To address these limitations, an automated self-configuring framework—analogous to nnU-Net [174]—could be envisioned, which might be termed "nn-ProtoNet". Building on ProtoPNet principles and integrating strengths from its existing variants, such a system would automatically optimize key design parameters (e.g., receptive fields, prototype numbers, training configurations) for a given dataset. This approach would lower the barrier to adoption, enhance robustness and generalization, and pave the way toward reliable, prototype-based models.

Hybrid CNN-Transformer models for self-explainable medical imaging. By reformulating classifier heads and attention components as convolutional operations, we extended inherent interpretability to hybrid CNN-Transformer architectures, enabling them to produce spatially resolved and faithful explanations without relying on post-hoc methods. The proposed hybrid CNN-Transformer achieved strong predictive performance on color fundus images for tasks such as diabetic retinopathy detection and age-related macular degeneration classification. Its framework naturally generalizes to other ophthalmic modalities, including retinal OCT for epiretinal membrane detection, as well as to broader clinical applications such as pneumonia detection from chest radiography and neurological disorder classification from brain imaging [16, 78, 79]. Future work could explore alternative CNN backbones to improve robustness across datasets and clinical contexts, while within the transformer component, architectural parameters such as attention window sizes, scaling factors in dual-resolution attention, and the number of pyramidal feature levels for multi-scale aggregation warrant systematic evaluation. Although the current implementation uses a single transformer block for computational efficiency, the convolution-based interpretability mechanism can be generalized to full-scale vision transformers, including ViTs [64], Swin [71], and ReTFound [175], a leading foundation model in ophthalmology. Extending this approach to large-scale foundation models will represent an important step toward clinically trustworthy AI, ensuring that high-performing vision transformer models also deliver transparent and reliable explanations.

Structured sparsity and smoothness in explanation maps. Sparse BagNet and SoftCAM leverage ElasticNet regularization to enhance interpretability by combining lasso penalties, which promote sparsity in explanations, with ridge penalties, which prevent omission of relevant disease features. Although this formulation effectively encourages *soft sparsity*, it does not explicitly account for the structured image patterns that are often needed in clinical contexts. For example, in Retinopathy of Prematurity classification [176], clinically meaningful explanation should exhibit both the sparsity and the spatial smoothness characteristic of retinal vessel—properties that ElasticNet cannot enforce. While the ℓ_1 norm promotes sparsity, it overlooks spatial continuity, potentially producing attribution maps misaligned with anatomical structures. A promising direction is to incorporate *hard sparsity* via ℓ_0 regularization, combined with explicit smoothness constraints, to generate explanations that more faithfully capture domain-specific structures [177]. Despite the inherent non-convexity of $\ell-_0$ regularization, optimization techniques such as the Alternating

Direction Method of Multipliers (ADMM) [178] offer practical solutions by reformulating it into a tractable convex optimization problem. Such a structured regularization could yield smoother and anatomically consistent explanations, thus improving interpretability.

Uncertainty and faithfulness in explainable models. Although the proposed interpretable models achieve classification performance comparable to black-box baselines while providing inherent explanations, true clinical applicability requires more than interpretability alone. In particular, it requires well-calibrated uncertainty estimates for both model predictions and their associated explanations [39,45]. However, deep neural networks are notoriously overconfident and lack inherently probabilistic foundations [179], compromising their reliability in safety-critical settings. Our proposed models—SoftCAM, sparse BagNet, Proto-BagNet, and the hybrid CNN-Transformer—could be further enhanced by integrating advanced probabilistic frameworks capable of quantifying uncertainty in both class-evidence maps and predicted labels [179,180]. This integration would allow clinicians not only to interpret the model's reasoning but also to assess its confidence, providing an additional safeguard against diagnostic errors. Beyond uncertainty estimation, our experiments revealed a fundamental challenge: the misalignment between *faithfulness metric*, which assesses how well explanations reflect the model's internal reasoning [181] and *top-k precision metrics* derived from the pointing game [156], which evaluate alignment with human or expert knowledge. Interestingly, the models that achieved the highest faithfulness scores were not always those whose explanations aligned best with expert annotations [5], consistent with earlier findings that highlight a systematic divergence between these evaluation dimensions [155]. Moreover, faithfulness-based metrics are not always sufficiently discriminative, as different explanation methods can lead to similar fidelity scores despite producing qualitatively distinct and clinically meaningful outputs [160]. These limitations underscore two key risks: relying solely on human-alignment metrics can overlook the model's actual reasoning process, while faithfulness metrics alone may fail to capture clinical interpretability. Therefore, we advocate for a multi-dimensional combined evaluation framework that jointly assesses both faithfulness and human alignment, ensuring that the explanations remain simultaneously faithful to the model's reasoning and practically useful to clinicians.

In conclusion, this thesis demonstrates that inherently self-explainable models can achieve competitive predictive performance while providing transparent, localized, and clinically meaningful explanations. By embedding interpretability directly into model architecture, these approaches address many of the shortcomings of post-hoc explanation methods, producing explanations that are faithful to the model's reasoning and actionable for end users. The findings emphasize the critical role of receptive field design, prototype diversity, convolutional layers, and structured sparsity in advancing interpretability for safety-critical applications. Together, these contributions establish a practical foundation for the design and rigorous evaluation of interpretable deep learning classifiers in medical imaging and beyond. They also outline clear future directions, including automated prototype optimization, prospective clinical validation, and the extension of inherent interpretability to large-scale foundation models—paving the way toward more trustworthy and deployable AI systems.

References

[1] Kerol Djoumessi, Indu Ilanchezian, Laura Kühlewein, Hanna Faber, Christian F Baumgartner, Bubacarr Bah, Philipp Berens, and Lisa M Koch. Sparse activations for interpretable disease grading. In *Medical Imaging with Deep Learning*, 2023.

[2] Kerol Djoumessi, Bubacarr Bah, Laura Kühlewein, Philipp Berens, and Lisa Koch. This actually looks like that: Proto-bagnets for local and global interpretability-by-design. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 718–728, 2024.

[3] Kerol Djoumessi, Ziwei Huang, Laura Kühlewein, Annekatrin Rickmann, Natalia Simon, Lisa M Koch, and Philipp Berens. An inherently interpretable ai model improves screening speed and accuracy for early diabetic retinopathy. *PLOS Digital Health*, 4(5):e0000831, 2025.

[4] Kerol Djoumessi, Samuel Ofosu Mensah, and Philipp Berens. A hybrid fully convolutional cnn-transformer model for inherently interpretable medical image classification. *arXiv preprint arXiv:2504.08481*, 2025.

[5] Kerol Djoumessi and Philipp Berens. Soft-cam: Making black box models self-explainable for high-stakes decisions. *arXiv preprint arXiv:2505.17748*, 2025.

[6] Xiaoxuan Liu, Livia Faes, Aditya U Kale, Siegfried K Wagner, Dun Jack Fu, Alice Bruynseels, Thushika Mahendiran, Gabriella Moraes, Mohith Shamdas, Christoph Kern, et al. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. *The lancet digital health*, 1(6):e271–e297, 2019.

[7] Xuemei Li, Alexander Sigov, Leonid Ratkin, Leonid A Ivanov, and Ling Li. Artificial intelligence applications in finance: a survey. *Journal of Management Analytics*, 10(4):676–692, 2023.

[8] Samira Pouyanfar, Saad Sadiq, Yilin Yan, Haiman Tian, Yudong Tao, Maria Presa Reyes, Mei-Ling Shyu, Shu-Ching Chen, and Sundaraja S Iyengar. A survey on deep learning: Algorithms, techniques, and applications. *ACM computing surveys (CSUR)*, 51(5):1–36, 2018.

[9] Shaveta Dargan, Munish Kumar, Maruthi Rohit Ayyagari, and Gulshan Kumar. A survey of deep learning and its applications: a new paradigm to machine learning. *Archives of computational methods in engineering*, 27(4):1071–1092, 2020.

[10] Kevin P Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012.

[11] Katarzyna Kolasa, Bisrat Admassu, Malwina Hołownia-Voloskova, Katarzyna J Kedzior, Jean-Etienne Poirrier, and Stefano Perni. Systematic reviews of machine learning in healthcare: a literature review. *Expert Review of Pharmacoeconomics & Outcomes Research*, 24(1):63–115, 2024.

[12] Urs J Muehlematter, Paola Daniore, and Kerstin N Vokinger. Approval of artificial intelligence and machine learning-based medical devices in the usa and europe (2015–20): a comparative analysis. *The Lancet Digital Health*, 3(3):e195–e203, 2021.

[13] Shi Dong, Ping Wang, and Khushnood Abbas. A survey on deep learning and its applications. *Computer Science Review*, 40:100379, 2021.

[14] Varun Gulshan, Lily Peng, Marc Coram, Martin C. Stumpe, Derek Wu, Arunachalam Narayanaswamy, Subhashini Venugopalan, Kasumi Widner, Tom Madams, Jorge Cuadros, Ramasamy Kim, Rajiv Raman, Philip C. Nelson, Jessica L. Mega, and Dale R. Webster. Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. *JAMA*, 316(22):2402, December 2016.

[15] Jakob Grauslund. Diabetic retinopathy screening in the emerging era of artificial intelligence. *Diabetologia*, 65(9):1415–1423, 2022.

[16] Eunji Kim, Siwon Kim, Minji Seo, and Sungroh Yoon. Xprotonet: diagnosis in chest radiography with global and local explanations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15719–15728, 2021.

[17] Amer Kareem, Haiming Liu, and Paul Sant. Review on pneumonia image detection: A machine learning approach. *Human-Centric Intelligent Systems*, 2(1):31–43, 2022.

[18] Milena Nacchia, Fabio Fruggiero, Alfredo Lambiase, and Ken Bruton. A systematic mapping of the advancing use of machine learning techniques for predictive maintenance in the manufacturing sector. *Applied Sciences*, 11(6):2546, 2021.

[19] Samreen Naeem, Aqib Ali, Sania Anam, and Muhammad Munawar Ahmed. An unsupervised machine learning algorithms: Comprehensive review. *International Journal of Computing and Digital Systems*, 2023.

[20] Richard S Sutton and Andrew G Barto. *Reinforcement Learning: An Introduction*. MIT Press, 2018.

[21] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen Awm Van Der Laak, Bram Van Ginneken, and Clara I Sánchez. A survey on deep learning in medical image analysis. *Medical image analysis*, 42:60–88, 2017.

[22] Philippe Bracke, Anupam Datta, Carsten Jung, and Shayak Sen. Machine learning explainability in finance: an application to default risk analysis. 2019.

[23] Eugene Gilmore, Vladimir Estivill-Castro, and René Hexel. More interpretable decision trees. In *Hybrid Artificial Intelligent Systems: 16th International Conference, HAIS 2021, Bilbao, Spain, September 22–24, 2021, Proceedings 16*, pages 280–292. Springer, 2021.

[24] Emina Tahirovic and Senka Krivic. Interpretability and explainability of logistic regression model for breast cancer detection. In *ICAART (3)*, pages 161–168, 2023.

[25] Khansa Rasheed, Adnan Qayyum, Mohammed Ghaly, Ala Al-Fuqaha, Adeel Razi, and Junaid Qadir. Explainable, trustworthy, and ethical machine learning for healthcare: A survey. *Computers in Biology and Medicine*, 149:106043, 2022.

[26] Bas HM Van der Velden, Hugo J Kuijf, Kenneth GA Gilhuijs, and Max A Viergever. Explainable artificial intelligence (xai) in deep learning-based medical image analysis. *Medical image analysis*, 79:102470, 2022.

[27] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*, 51(5):1–42, 2018.

[28] Hassija, Vinay Chamola, Atmesh Mahapatra, Abhinandan Singal, Divyansh Goel, Kaizhu Huang, Simone Scardapane, Indro Spinelli, Mufti Mahmud, and Amir Hussain. Interpreting black-box models: a review on explainable artificial intelligence. *Cognitive Computation*, 16(1):45–74, 2024.

[29] Rabia Saleem, Bo Yuan, Fatih Kurugollu, Ashiq Anjum, and Lu Liu. Explaining deep neural networks: A survey on the global interpretation methods. *Neurocomputing*, 513:165–180, 2022.

[30] Vikas Hassija, Balamurugan Palanisamy, Arpita Chatterjee, Arpita Mandal, Debanshi Chakraborty, Amit Pandey, GSS Chalapathi, and Dhruv Kumar. Transformers for vision: A survey on innovative methods for computer vision. *IEEE Access*, 2025.

[31] José Maurício, Inês Domingues, and Jorge Bernardino. Comparing vision transformers and convolutional neural networks for image classification: A literature review. *Applied Sciences*, 13(9):5521, 2023.

[32] Satoshi Takahashi, Yusuke Sakaguchi, Nobuji Kouno, Ken Takasawa, Kenichi Ishizu, Yu Akagi, Rina Aoyama, Naoki Teraya, Amina Bolatkan, Norio Shinkai, et al. Comparison of vision transformers and convolutional neural networks in medical image analysis: A systematic review. *Journal of Medical Systems*, 48(1):1–22, 2024.

[33] Cynthia Rudin. Why black box machine learning should be avoided for high-stakes decisions, in brief. *Nature Reviews Methods Primers*, 2(1):81, 2022.

[34] 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.

[35] Amina Adadi and Mohammed Berrada. Explainable ai for healthcare: from black box to interpretable models. In *Embedded systems and artificial intelligence: proceedings of ESAI 2019, Fez, Morocco*, pages 327–337. Springer, 2020.

[36] Yu Zhang, Peter Tiňo, Aleš Leonardis, and Ke Tang. A survey on neural network interpretability. *IEEE transactions on emerging topics in computational intelligence*, 5(5):726–742, 2021.

[37] Geeta Joshi, Aditi Jain, Shalini Reddy Araveeti, Sabina Adhikari, Harshit Garg, and Mukund Bhandari. Fda-approved artificial intelligence and machine learning (ai/ml)-enabled medical devices: an updated landscape. *Electronics*, 13(3):498, 2024.

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

[39] Protection Regulation. Regulation (eu) 2016/679 of the european parliament and of the council. *Regulation (eu)*, 679(2016):10–13, 2016.

[40] Christoph Molnar. *Interpretable machine learning*. Lulu. com, 2020.

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

[42] Julius Adebayo, Michael Muelly, Ilaria Liccardi, and Been Kim. Debugging tests for model explanations. *Advances in Neural Information Processing Systems*, 33:700–712, 2020.

[43] 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.

[44] Luca Nannini, Agathe Balayn, and Adam Leon Smith. Explainability in ai policies: A critical review of communications, reports, regulations, and standards in the eu, us, and uk. In *Proceedings of the 2023 ACM conference on fairness, accountability, and transparency*, pages 1198–1212, 2023.

[45] Nathalie A Smuha. Regulation 2024/1689 of the eur. parl. & council of june 13, 2024 (eu artificial intelligence act). *International Legal Materials*, pages 1–148, 2025.

[46] Paul Voigt and Axel Von dem Bussche. The eu general data protection regulation (gdpr). *A practical guide, 1st ed., Cham: Springer International Publishing*, 10(3152676):10–5555, 2017.

[47] 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.

[48] Georgios Pavlidis. Unlocking the black box: analysing the eu artificial intelligence act's framework for explainability in ai. *Law, Innovation and Technology*, 16(1):293–308, 2024.

[49] Cristina Gonzalez-Gonzalo, Eric F Thee, Caroline CW Klaver, Aaron Y Lee, Reinier O Schlingemann, Adnan Tufail, Frank Verbraak, and Clara I Sánchez. Trustworthy ai: Closing the gap between development and integration of ai systems in ophthalmic practice. *Progress in retinal and eye research*, 90:101034, 2022.

[50] Thomas Grote and Philipp Berens. How competitors become collaborators—bridging the gap (s) between machine learning algorithms and clinicians. *Bioethics*, 36(2):134–142, 2022.

[51] Thomas Grote and Philipp Berens. On the ethics of algorithmic decision-making in healthcare. *Journal of medical ethics*, 46(3):205–211, 2020.

[52] Ms Aayushi Bansal, Dr Rewa Sharma, and Dr Mamta Kathuria. A systematic review on data scarcity problem in deep learning: solution and applications. *ACM Computing Surveys (Csur)*, 54(10s):1–29, 2022.

[53] Aida Ali, Siti Mariyam Shamsuddin, and Anca L Ralescu. Classification with class imbalance problem. *Int. J. Advance Soft Compu. Appl*, 5(3):176–204, 2013.

[54] Jingkang Yang, Kaiyang Zhou, Yixuan Li, and Ziwei Liu. Generalized out-of-distribution detection: A survey. *International Journal of Computer Vision*, 132(12):5635–5662, 2024.

[55] Lisa M Koch, Christian F Baumgartner, and Philipp Berens. Distribution shift detection for the postmarket surveillance of medical ai algorithms: a retrospective simulation study. *NPJ Digital Medicine*, 7(1):120, 2024.

[56] Wieland Brendel and Matthias Bethge. Approximating cnns with bag-of-local-features models works surprisingly well on imagenet. *International Conference on Learning Representations*, 2019.

[57] 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.

[58] Julius Gervelmeyer, Sarah Müller, Kerol Djoumessi, David Merle, Simon J Clark, Lisa Koch, and Philipp Berens. Interpretable-by-design deep survival analysis for disease progression modeling. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 502–512. Springer, 2024.

[59] Samuel Ofosu Mensah, Jonas Neubauer, Murat Seckin Ayhan, Kerol Roussin Djoumessi Donteu, Lisa Margaret Koch, Murat Mehmet Uzel, Faik Gelisken, and Philipp Berens. Clinically interpretable deep learning via sparse bagnets for epiretinal membrane and related pathology detection. *medRxiv*, pages 2025–06, 2025.

[60] Samuel Ofosu Mensah, Kerol Djoumessi, and Philipp Berens. Prototype-guided and lightweight adapters for inherent interpretation and generalisation in federated learning. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 464–473. Springer, 2025.

[61] Léon Bottou et al. Stochastic gradient learning in neural networks. *Proceedings of Neuro-Nımes*, 91(8):12, 1991.

[62] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning internal representations by error propagation. Technical report, 1985.

[63] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[64] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR*, 2021.

[65] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[66] Zewen Li, Fan Liu, Wenjie Yang, Shouheng Peng, and Jun Zhou. A survey of convolutional neural networks: analysis, applications, and prospects. *IEEE transactions on neural networks and learning systems*, 33(12):6999–7019, 2021.

[67] Reza Azad, Amirhossein Kazerouni, Moein Heidari, Ehsan Khodapanah Aghdam, Amirali Molaei, Yiwei Jia, Abin Jose, Rijo Roy, and Dorit Merhof. Advances in medical image analysis with vision transformers: a comprehensive review. *Medical Image Analysis*, 91:103000, 2024.

[68] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

[69] Sepp Hochreiter. The vanishing gradient problem during learning recurrent neural nets and problem solutions. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 6(02):107–116, 1998.

[70] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30, 2017.

[71] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.

[72] Kushal Chauhan, Rishabh Tiwari, Jan Freyberg, Pradeep Shenoy, and Krishnamurthy Dvijotham. Interactive concept bottleneck models. In *Proceedings of the aaai conference on artificial intelligence*, volume 37, pages 5948–5955, 2023.

[73] Alina Jade Barnett, Fides Regina Schwartz, Chaofan Tao, Chaofan Chen, Yinhao Ren, Joseph Y Lo, and Cynthia Rudin. A case-based interpretable deep learning model for classification of mass lesions in digital mammography. *Nature Machine Intelligence*, 3(12):1061–1070, 2021.

[74] Rory Sayres, Ankur Taly, Ehsan Rahimy, Katy Blumer, David Coz, Naama Hammel, Jonathan Krause, Arunachalam Narayanaswamy, Zahra Rastegar, Derek Wu, et al. Using a deep learning algorithm and integrated gradients explanation to assist grading for diabetic retinopathy. *Ophthalmology*, 126(4):552–564, 2019.

[75] Gwilym S Lodwick. Computer-aided diagnosis in radiology: A research plan. *Investigative Radiology*, 1(1):72–80, 1966.

[76] Jun-Ichiro Toriwaki, Yasuhito Suenaga, Toshio Negoro, and Teruo Fukumura. Pattern recognition of chest x-ray images. *Computer Graphics and Image Processing*, 2(3-4):252–271, 1973.

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

[78] Tahir Hussain, Hayaru Shouno, Abid Hussain, Dostdar Hussain, Muhammad Ismail, Tatheer Hussain Mir, Fang Rong Hsu, Taukir Alam, and Shabnur Anonna Akhy. Effresnet-vit: A fusion-based convolutional and vision transformer model for explainable medical image classification. *IEEE Access*, 2025.

[79] George Shih, Carol C Wu, Safwan S Halabi, Marc D Kohli, Luciano M Prevedello, Tessa S Cook, Arjun Sharma, Judith K Amorosa, Veronica Arteaga, Maya Galperin-Aizenberg, et al. Augmenting the national institutes of health chest radiograph dataset with expert annotations of possible pneumonia. *Radiology: Artificial Intelligence*, 1(1):e180041, 2019.

[80] Emma Dugas, Jorge Jared, and Will Cukierski. Diabetic retinopathy detection, 2015.

[81] Daniel S Kermany, Michael Goldbaum, Wenjia Cai, Carolina CS Valentim, Huiying Liang, Sally L Baxter, Alex McKeown, Ge Yang, Xiaokang Wu, Fangbing Yan, et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. *cell*, 172(5):1122–1131, 2018.

[82] Michael David Abràmoff, Yiyue Lou, Ali Erginay, Warren Clarida, Ryan Amelon, James C Folk, and Meindert Niemeijer. Improved automated detection of diabetic retinopathy on a publicly available dataset through integration of deep learning. *Investigative ophthalmology & visual science*, 57(13):5200–5206, 2016.

[83] Jorge Cuadros and George Bresnick. Eyepacs: an adaptable telemedicine system for diabetic retinopathy screening. *Journal of diabetes science and technology*, 3(3):509–516, 2009.

[84] Tien Y Wong, Jennifer Sun, Ryo Kawasaki, Paisan Ruamviboonsuk, Neeru Gupta, Van Charles Lansingh, Mauricio Maia, Wanjiku Mathenge, Sunil Moreker, Mahi MK Muqit, et al. Guidelines on diabetic eye care: the international council of ophthalmology recommendations for screening, follow-up, referral, and treatment based on resource settings. *Ophthalmology*, 125(10):1608–1622, 2018.

[85] M Kiruthika and G Malathi. A comprehensive review on early detection of drusen patterns in age-related macular degeneration using deep learning models. *Photodiagnosis and Photodynamic Therapy*, page 104454, 2024.

[86] Qingsheng Peng, Rachel Marjorie Wei Wen Tseng, Yih-Chung Tham, Ching-Yu Cheng, and Tyler Hyungtaek Rim. Detection of systemic diseases from ocular images using artificial intelligence: a systematic review. *The Asia-Pacific Journal of Ophthalmology*, 11(2):126–139, 2022.

[87] Erdi Çallı, Ecem Sogancioglu, Bram Van Ginneken, Kicky G van Leeuwen, and Keelin Murphy. Deep learning for chest x-ray analysis: A survey. *Medical image analysis*, 72:102125, 2021.

[88] Jakob Weiss, Vineet K Raghu, Kaavya Paruchuri, Aniket Zinzuwadia, Pradeep Natarajan, Hugo JWL Aerts, and Michael T Lu. Deep learning to estimate cardiovascular risk from chest radiographs: a risk prediction study. *Annals of internal medicine*, 177(4):409–417, 2024.

[89] Chris Solomou and Dimitar Kazakov. Utilizing chest x-rays for age prediction and gender classification. In *2021 4th International Seminar on Research of Information Technology and Intelligent Systems (ISRITI)*, pages 356–361. IEEE, 2021.

[90] Indu Ilanchezian. *Visually Explaining Decisions of Deep Neural Network Classifiers in Ophthalmology*. PhD thesis, Dissertation, Tübingen, Universität Tübingen, 2024, 2024.

[91] Kyriakos D Apostolidis and George A Papakostas. A survey on adversarial deep learning robustness in medical image analysis. *Electronics*, 10(17):2132, 2021.

[92] Weimin Zhao, Sanaa Alwidian, and Qusay H Mahmoud. Adversarial training methods for deep learning: A systematic review. *Algorithms*, 15(8):283, 2022.

[93] Ruinan Jin, Zikang Xu, Yuan Zhong, Qingsong Yao, DOU QI, S Kevin Zhou, and Xiaoxiao Li. Fairmedfm: fairness benchmarking for medical imaging foundation models. *Advances in Neural Information Processing Systems*, 37:111318–111357, 2024.

[94] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)*, 54(6):1–35, 2021.

[95] Gaurav Menghani. Efficient deep learning: A survey on making deep learning models smaller, faster, and better. *ACM Computing Surveys*, 55(12):1–37, 2023.

[96] Suruchi Kumari and Pravendra Singh. Data efficient deep learning for medical image analysis: A survey. *arXiv preprint arXiv:2310.06557*, 2023.

[97] Defu Liu, Yixiao Zhu, Zhe Liu, Yi Liu, Changlin Han, Jinkai Tian, Ruihao Li, and Wei Yi. A survey of model compression techniques: Past, present, and future. *Frontiers in Robotics and AI*, 12:1518965, 2025.

[98] Mateusz Cedro and David Martens. Graphxain: Narratives to explain graph neural networks. *arXiv preprint arXiv:2411.02540*, 2024.

[99] Jiaxing Zhang, Jiayi Liu, Dongsheng Luo, Jennifer Neville, and Hua Wei. Llmexplainer: Large language model based bayesian inference for graph explanation generation. *arXiv preprint arXiv:2407.15351*, 2024.

[100] Leo Breiman. Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical science*, 16(3):199–231, 2001.

[101] Philipp Tschandl, Christoph Rinner, Zoe Apalla, Giuseppe Argenziano, Noel Codella, Allan Halpern, Monika Janda, Aimilios Lallas, Caterina Longo, Josep Malvehy, et al. Human–computer collaboration for skin cancer recognition. *Nature medicine*, 26(8):1229–1234, 2020.

[102] Pascal Sturmfels, Scott Lundberg, and Su-In Lee. Visualizing the impact of feature attribution baselines. *Distill*, 5(1):e22, 2020.

[103] Joseph D Janizek, Ayse B Dincer, Safiye Celik, Hugh Chen, William Chen, Kamila Naxerova, and Su-In Lee. Uncovering expression signatures of synergistic drug responses via ensembles of explainable machine-learning models. *Nature biomedical engineering*, 7(6):811–829, 2023.

[104] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer, 2014.

[105] Matthieu Bellucci, Nicolas Delestre, Nicolas Malandain, and Cecilia Zanni-Merk. Towards a terminology for a fully contextualized xai. *Procedia Computer Science*, 192:241–250, 2021.

[106] Junlin Hou, Sicen Liu, Yequan Bie, Hongmei Wang, Andong Tan, Luyang Luo, and Hao Chen. Self-explainable ai for medical image analysis: A survey and new outlooks. *arXiv preprint arXiv:2410.02331*, 2024.

[107] Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *International conference on machine learning*, pages 5338–5348. PMLR, 2020.

[108] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[109] Peng-Tao Jiang, Chang-Bin Zhang, Qibin Hou, Ming-Ming Cheng, and Yunchao Wei. Layercam: Exploring hierarchical class activation maps for localization. *IEEE Transactions on Image Processing*, 2021.

[110] Scott M Lundberg, Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee. From local explanations to global understanding with explainable ai for trees. *Nature machine intelligence*, 2(1):56–67, 2020.

[111] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016.

[112] Haofan Wang, Zifan Wang, Mengnan Du, Fan Yang, Zijian Zhang, Sirui Ding, Piotr Mardziel, and Xia Hu. Score-cam: Score-weighted visual explanations for convolutional neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 24–25, 2020.

[113] Sahra Ghalebikesabi, Lucile Ter-Minassian, Karla DiazOrdaz, and Chris C Holmes. On locality of local explanation models. *Advances in neural information processing systems*, 34:18395–18407, 2021.

[114] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

[115] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

[116] Vitali Petsiuk, Abir Das, and Kate Saenko. Rise: Randomized input sampling for explanation of black-box models. *arXiv preprint arXiv:1806.07421*, 2018.

[117] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR, 2017.

[118] Eleonora Poeta, Gabriele Ciravegna, Eliana Pastor, Tania Cerquitelli, and Elena Baralis. Concept-based explainable artificial intelligence: A survey. *CoRR*, 2023.

[119] Amirata Ghorbani, James Wexler, James Y Zou, and Been Kim. Towards automatic concept-based explanations. *Advances in neural information processing systems*, 32, 2019.

[120] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International conference on machine learning*, pages 2668–2677. PMLR, 2018.

[121] Nicolas Papernot and Patrick McDaniel. Deep k-nearest neighbors: Towards confident, interpretable and robust deep learning. *arXiv preprint arXiv:1803.04765*, 2018.

[122] Mingwei He, Bohan Li, and Songlin Sun. A survey of class activation mapping for the interpretability of convolution neural networks. In *International Conference On Signal And Information Processing, Networking And Computers*, pages 399–407. Springer, 2022.

[123] Marco Ancona, Enea Ceolini, Cengiz Öztireli, and Markus Gross. Towards better understanding of gradient-based attribution methods for deep neural networks. *arXiv preprint arXiv:1711.06104*, 2017.

[124] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. Striving for simplicity: The all convolutional net. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Workshop Track Proceedings*, 2014.

[125] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7), 2015.

[126] Marko Robnik-Šikonja and Marko Bohanec. Perturbation-based explanations of prediction models. In *Human and Machine Learning: Visible, Explainable, Trustworthy and Transparent*, pages 159–175. Springer, 2018.

[127] Anuja Vats, Marius Pedersen, and Ahmed Mohammed. Concept-based reasoning in medical imaging. *International Journal of Computer Assisted Radiology and Surgery*, 18(7):1335–1339, 2023.

[128] Chih-Kuan Yeh, Been Kim, Sercan Arik, Chun-Liang Li, Tomas Pfister, and Pradeep Ravikumar. On completeness-aware concept-based explanations in deep neural networks. *Advances in neural information processing systems*, 33:20554–20565, 2020.

[129] Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *International conference on machine learning*, pages 1885–1894. PMLR, 2017.

[130] Been Kim, Rajiv Khanna, and Oluwasanmi O Koyejo. Examples are not enough, learn to criticize! criticism for interpretability. *Advances in neural information processing systems*, 29, 2016.

[131] Riccardo Guidotti. Counterfactual explanations and how to find them: literature review and benchmarking. *Data Mining and Knowledge Discovery*, 38(5):2770–2824, 2024.

[132] Nishanth Arun, Nathan Gaw, Praveer Singh, Ken Chang, Mehak Aggarwal, Bryan Chen, Katharina Hoebel, Sharut Gupta, Jay Patel, Mishka Gidwani, et al. Assessing the trustworthiness of saliency maps for localizing abnormalities in medical imaging. *Radiology: Artificial Intelligence*, 3(6):e200267, 2021.

[133] Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5):359–366, 1989.

[134] Lisa Anita De Santi, Franco Italo Piparo, Filippo Bargagna, Maria Filomena Santarelli, Simona Celi, and Vincenzo Positano. Part-prototype models in medical imaging: Applications and current challenges. *BioMedInformatics*, 4(4):2149–2172, 2024.

[135] Jian Cao, Dian-hui Mao, Qiang Cai, Hai-sheng Li, and Jun-ping Du. A review of object representation based on local features. *Journal of Zhejiang University SCIENCE C*, 14(7):495–504, 2013.

[136] Khawla Elhadri, Tomasz Michalski, Adam Wróbel, Jörg Schlötterer, Bartosz Zieliński, and Christin Seifert. This looks like what? challenges and future research directions for part-prototype models. *arXiv preprint arXiv:2502.09340*, 2025.

[137] 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.

[138] Luisa Gallée, Meinrad Beer, and Michael Götz. Interpretable medical image classification using prototype learning and privileged information. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 435–445. Springer, 2023.

[139] Yuanyuan Wei, Roger Tam, and Xiaoying Tang. Mprotonet: A case-based interpretable model for brain tumor classification with 3d multi-parametric magnetic resonance imaging. In *Medical Imaging with Deep Learning*, pages 1798–1812. PMLR, 2024.

[140] Poulami Sinhamahapatra, Lena Heidemann, Maureen Monnet, and Karsten Roscher. Towards human-interpretable prototypes for visual assessment of image classification models. *arXiv preprint arXiv:2211.12173*, 2022.

[141] Romain Xu-Darme, Georges Quénot, Zakaria Chihani, and Marie-Christine Rousset. Sanity checks for patch visualisation in prototype-based image classification. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023 - Workshops, Vancouver, BC, Canada, June 17-24, 2023*. IEEE, 2023.

[142] Srishti Gautam, Marina M-C Höhne, Stine Hansen, Robert Jenssen, and Michael Kampffmeyer. This looks more like that: Enhancing self-explaining models by prototypical relevance propagation. *Pattern Recognition*, 136:109172, 2023.

[143] Prasanna Porwal, Samiksha Pachade, Ravi Kamble, Manesh Kokare, Girish Deshmukh, Vivek Sahasrabuddhe, and Fabrice Meriaudeau. Indian diabetic retinopathy image dataset (idrid): a database for diabetic retinopathy screening research. *Data*, 3(3):25, 2018.

[144] Peng Li, Cunqian Feng, Xiaowei Hu, and Zixiang Tang. Sar-bagnet: An ante-hoc interpretable recognition model based on deep network for sar image. *Remote Sensing*, 14(9):2150, 2022.

[145] Dichao Liu and Kenji Suzuki. Edge-and color–texture-aware bag-of-local-features model for accurate and interpretable skin lesion diagnosis. *Diagnostics*, 15(15):1883, 2025.

[146] Gergo Galiger and Z Bodo. Explainable patch-level histopathology tissue type detection with bag-of-local-features models and data augmentation. *ACTA Univ. Sapientiae Inform*, 15:60–80, 2023.

[147] Sheng He, P Ellen Grant, and Yangming Ou. Global-local transformer for brain age estimation. *IEEE transactions on medical imaging*, 41(1):213–224, 2021.

[148] Moritz Bohle, Mario Fritz, and Bernt Schiele. Convolutional dynamic alignment networks for interpretable classifications. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10029–10038, 2021.

[149] Moritz Böhle, Mario Fritz, and Bernt Schiele. B-cos networks: Alignment is all we need for interpretability. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10329–10338, 2022.

[150] Jianlong Zhou, Amir H Gandomi, Fang Chen, and Andreas Holzinger. Evaluating the quality of machine learning explanations: A survey on methods and metrics. *Electronics*, 10(5):593, 2021.

[151] Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *stat*, 1050:2, 2017.

[152] Weina Jin, Xiaoxiao Li, Mostafa Fatehi, and Ghassan Hamarneh. Guidelines and evaluation of clinical explainable ai in medical image analysis. *Medical image analysis*, 84:102684, 2023.

[153] 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.

[154] 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: A survey of user studies for model explanations. *IEEE transactions on pattern analysis and machine intelligence*, 46(4):2104–2122, 2023.

[155] Chih-Kuan Yeh, Cheng-Yu Hsieh, Arun Suggala, David I Inouye, and Pradeep K Ravikumar. On the (in) fidelity and sensitivity of explanations. *Advances in neural information processing systems*, 32, 2019.

[156] Jianming Zhang, Sarah Adel Bargal, Zhe Lin, Jonathan Brandt, Xiaohui Shen, and Stan Sclaroff. Top-down neural attention by excitation backprop. *International Journal of Computer Vision*, 126(10):1084–1102, 2018.

[157] Meike Nauta and Christin Seifert. The co-12 recipe for evaluating interpretable part-prototype image classifiers. In *World Conference on Explainable Artificial Intelligence*, pages 397–420. Springer, 2023.

[158] Qihan Huang, Mengqi Xue, Wenqi Huang, Haofei Zhang, Jie Song, Yongcheng Jing, and Mingli Song. Evaluation and improvement of interpretability for self-explainable part-prototype networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2011–2020, 2023.

[159] Tania Lombrozo. 14 explanation and abductive inference. *The Oxford handbook of thinking and reasoning*, page 260, 2012.

[160] Hanwei Zhang, Felipe Torres Figueroa, and Holger Hermanns. Saliency maps give a false sense of explainability to image classifiers: An empirical evaluation across methods and metrics. In *The 16th Asian Conference on Machine Learning (Conference Track)*, 2024.

[161] Ombretta Strafforello, Xin Liu, Klamer Schutte, and Jan van Gemert. Video bagnet: short temporal receptive fields increase robustness in long-term action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 159–166, 2023.

[162] Murat Seçkin Ayhan, Louis Benedikt Kümmerle, Laura Kühlewein, Werner Inhoffen, Gulnar Aliyeva, Focke Ziemssen, and Philipp Berens. Clinical validation of saliency maps for understanding deep neural networks in ophthalmology. *Medical Image Analysis*, 77:102364, 2022.

[163] Tao Li, Yingqi Gao, Kai Wang, Song Guo, Hanruo Liu, and Hong Kang. Diagnostic assessment of deep learning algorithms for diabetic retinopathy screening. *Information Sciences*, 501:511–522, 2019.

[164] Zaid Ilyas, Afsah Saleem, David Suter, John T Schousboe, William D Leslie, Joshua R Lewis, and Syed Zulqarnain Gilani. A hybrid cnn-transformer feature pyramid network for granular abdominal aortic calcification detection from dxa images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 14–25. Springer, 2024.

[165] Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4190–4197, July 2020.

[166] Hila Chefer, Shir Gur, and Lior Wolf. Transformer interpretability beyond attention visualization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 782–791, 2021.

[167] Age-Related Eye Disease Study Research Group. The age-related eye disease study (areds): Design implications areds report no. 1. *Controlled clinical trials*, 20(6):573–600, 1999.

[168] Haiping Wu, Bin Xiao, Noel Codella, Mengchen Liu, Xiyang Dai, Lu Yuan, and Lei Zhang. Cvt: Introducing convolutions to vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 22–31, 2021.

[169] Eduard Hogea, Darian M Onchis, Ana Coporan, Adina Magda Florea, and Codruta Istin. Vitmix: Vision transformer explainability augmented by mixed visualization methods. *arXiv preprint arXiv:2412.14231*, 2024.

[170] Chiyu Ma, Jon Donnelly, Wenjun Liu, Soroush Vosoughi, Cynthia Rudin, and Chaofan Chen. Interpretable image classification with adaptive prototype-based vision transformers. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[171] David R Cox. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202, 1972.

[172] Maxime Oquab, Léon Bottou, Ivan Laptev, and Josef Sivic. Is object localization for free?-weakly-supervised learning with convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 685–694, 2015.

[173] Omid Davoodi, Shayan Mohammadizadehsamakosh, and Majid Komeili. On the interpretability of part-prototype based classifiers: a human centric analysis. *Scientific Reports*, 13(1):23088, 2023.

[174] Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods*, 18(2):203–211, 2021.

[175] Yukun Zhou, Mark A Chia, Siegfried K Wagner, Murat S Ayhan, Dominic J Williamson, Robbert R Struyven, Timing Liu, Moucheng Xu, Mateo G Lozano, Peter Woodward-Court, et al. A foundation model for generalizable disease detection from retinal images. *Nature*, 622(7981):156–163, 2023.

[176] Eun Hee Hong, Yong Un Shin, and Heeyoon Cho. Retinopathy of prematurity: a review of epidemiology and current treatment strategies. *Clinical and Experimental Pediatrics*, 65(3):115, 2021.

[177] Ruth Fong, Mandela Patrick, and Andrea Vedaldi. Understanding deep networks via extremal perturbations and smooth masks. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2950–2958, 2019.

[178] Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, Jonathan Eckstein, et al. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine learning*, 3(1):1–122, 2011.

[179] Mohsin Akram, Muhammad Adnan, Syed Farooq Ali, Jameel Ahmad, Amr Yousef, Tagrid Abdullah N Alshalali, and Zaffar Ahmed Shaikh. Uncertainty-aware diabetic retinopathy detection using deep learning enhanced by bayesian approaches. *Scientific Reports*, 15(1):1342, 2025.

[180] Murat Seçkin Ayhan, Laura Kühlewein, Gulnar Aliyeva, Werner Inhoffen, Focke Ziemssen, and Philipp Berens. Expert-validated estimation of diagnostic uncertainty for deep neural networks in diabetic retinopathy detection. *Medical image analysis*, 64:101724, 2020.

[181] David Alvarez-Melis and Tommi S Jaakkola. On the robustness of interpretability methods. *arXiv preprint arXiv:1806.08049*, 2018.

A Complete publications

This thesis is based on the following publications. An overview of contributions is provided in Chapter 3, within the corresponding paper section. The full quantitative contribution details are provided in a separate document alongside the thesis.

1. **Kerol Djoumessi**, Indu Ilanchezian, Laura Kühlewein, Hanna Faber, Christian F Baumgartner, Bubacarr Bah, Philipp Berens, and Lisa Koch. "Sparse activations for interpretable disease grading". In *Medical Imaging with Deep Learning*, 2023.

2. **Kerol Djoumessi**, Ziwei Huang, Laura Kühlewein, Annekatrin Rickmann, Natalia Simon, Lisa Koch, and Philipp Berens. "An inherently interpretable AI model improves screening speed and accuracy for early diabetic retinopathy". *PLOS Digital Health*, 2025.

3. **Kerol Djoumessi**, Bubacarr Bah, Laura Kühlewein, Philipp Berens, and Lisa Koch. "This actually looks like that: Proto-bagnets for local and global interpretability-by-design". In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 718–728, 2024.

4. **Kerol Djoumessi** and Philipp Berens. "Soft-CAM: Making black box models self-explainable for high-stakes decisions". *arXiv preprint arXiv:2505.17748*, 2025.

5. **Kerol Djoumessi**, Samuel Mensah, and Philipp Berens. "A Hybrid Fully Convolutional CNN-Transformer Model for Inherently Interpretable Disease Detection from Retinal Fundus Images". Accepted and presented at the *International Workshop on Interpretability of Machine Intelligence in Medical Image Computing (IMIMIC)*, 2025.

Sparse Activations for Interpretable Disease Grading

Kerol R. Donteu Djoumessi[1,2] KEROL.DJOUMESSI@UNI-TUEBINGEN.DE

Indu Ilanchezian[1,2] INDU.ILANCHEZIAN@UNI-TUEBINGEN.DE

Laura Kühlewein[3] LAURA.KUEHLEWEIN@MED.UNI-TUEBINGEN.DE

Hanna Faber[3] HANNA.FABER@MED.UNI-TUEBINGEN.DE

Christian F. Baumgartner[4] CHRISTIAN.BAUMGARTNER@UNI-TUEBINGEN.DE

Bubacarr Bah[5,6] BUBACARR@AIMS.AC.ZA

Philipp Berens[1,2,7] PHILIPP.BERENS@UNI-TUEBINGEN.DE

Lisa M. Koch[1,2] LISA.KOCH@UNI-TUEBINGEN.DE

[1] *Hertie Institute for Artificial Intelligence in Brain Health, University of Tübingen, Germany*
[2] *Institute of Ophthalmic Research, University of Tübingen, Germany*
[3] *University Eye Clinic, University of Tübingen, Germany*
[4] *Department of Computer Science, University of Tübingen, Germany*
[5] *African Institute for Mathematical Sciences (AIMS) South Africa and Stellenbosh University*
[6] *Medical Research Council Unit The Gambia at London School of Hygiene and Tropical Medicine*
[7] *Tübingen AI Center, University of Tübingen, Germany*

Abstract

Interpreting deep learning models typically relies on post-hoc saliency map techniques. However, these techniques often fail to serve as actionable feedback to clinicians, and they do not directly explain the decision mechanism. Here, we propose an inherently interpretable model that combines the feature extraction capabilities of deep neural networks with advantages of sparse linear models in interpretability. Our approach relies on straightforward but effective changes to a deep bag-of-local-features model (BagNet). These modifications lead to fine-grained and sparse class evidence maps which, by design, correctly reflect the model's decision mechanism. Our model is particularly suited for tasks which rely on characterising regions of interests that are very small and distributed over the image. In this paper, we focus on the detection of Diabetic Retinopathy, which is characterised by the progressive presence of small retinal lesions on fundus images. We observed good classification accuracy despite our added sparseness constraint. In addition, our model precisely highlighted retinal lesions relevant for the disease grading task and excluded irrelevant regions from the decision mechanism. The results suggest our sparse BagNet model can be a useful tool for clinicians as it allows efficient inspection of the model predictions and facilitates clinicians' and patients' trust.

Keywords: interpretability, sparse activations, diabetic retinopathy

1. Introduction

While machine learning (ML) tools have been approaching expert-level performance in many medical imaging tasks thanks to progress in deep learning (De Fauw et al., 2018; Shen et al., 2019; Mahoro and Akhloufi, 2022), they lack interpretability thereby posing ethical concerns (Grote and Berens, 2020) and preventing wide adoption in clinical practice (Teng et al., 2022). Deep ML models are most commonly explained by identifying image regions that influence the output of the trained model with post-hoc saliency maps (Simonyan et al., 2013; Zhou et al., 2016; Springenberg et al., 2014; Selvaraju et al., 2020). However, using saliency maps has been recently shown to be problematic for medical images as they only poorly localize disease-related lesions and are highly variable (Arun et al., 2021; Saporta et al., 2022). Furthermore, they do not provide actionable insights, given that they do not directly reflect the network's actual decision mechanisms. Instead, inherently interpretable models could provide a path forward for safety-critical tasks (Rudin, 2019), but few such models achieve sufficiently high prediction accuracy at the same time. In particular, classical linear models perform poorly when directly applied to medical images.

In this paper, we develop an inherently interpretable deep learning model that combines the feature extraction capabilities of deep neural networks with the advantages in interpretability of sparse linear models. Our model is especially suited for clinically relevant tasks which require identifying and characterising small lesions or other anomalies in large search regions. Examples of such tasks include screening for certain retinal diseases, breast or lung cancer. We focus here on the detection and grading of Diabetic Retinopathy (DR) on retinal fundus images.

DR is a microvascular complication of diabetes characterized by the progressive presence of one or more small retinal lesions such as microaneurysms, hemorrhages, or hard and soft exudates (ICO, 2017). It is the leading cause of blindness in the working-age population and the third leading cause of visual impairment worldwide, and early diagnosis and treatment can slow its progression (ICO, 2017; Wong et al., 2018). It is therefore recommended that diabetes patients undergo regular monitoring, and ML could facilitate mass screening and help clinicians use their time more efficiently (Ting et al., 2016).

Numerous high-performing black-box DR detection methods have been proposed (Rao et al., 2020; Alyoubi et al., 2020; Tavakoli and Kelley, 2021; Huang et al., 2021). For such methods, interpretation is mostly aided by saliency maps (Wang and Yang, 2019; Chetoui and Akhloufi, 2020) or the generation of counterfactual images (González-Gonzalo et al., 2020; Boreiko et al., 2022). A more interpretable approach for detecting DR is a multiple-instance learning model which combines features extracted from different image patches with attention weights (Papadopoulos et al., 2021). These weights can be visualised as a heatmap showing the contribution of different image regions to the prediction. Although all these methods provide some visual evidence of suspicious image regions, the saliency maps do not directly explain the decision mechanism, making their interpretation unintuitive. Further, they are often too cluttered and dense to be useful as feedback for clinicians and may be too coarse to identify small lesions.

We overcome these key limitations and propose a model for DR detection and grading which performs comparably to state-of-the-art models, despite being interpretable-by-

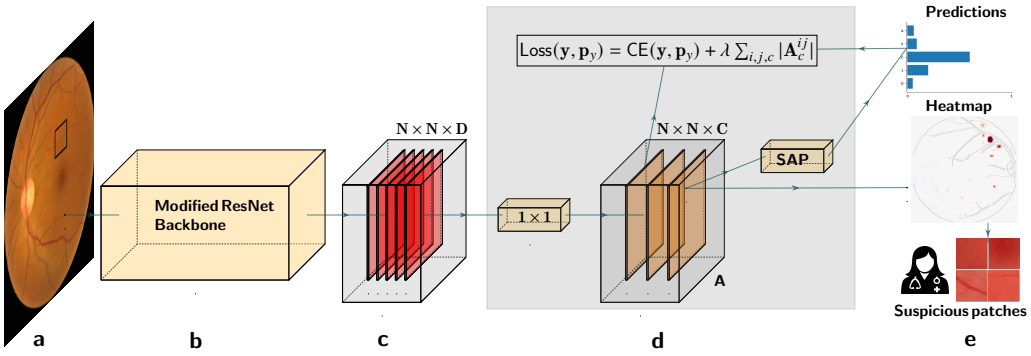

Figure 1: Overview of the proposed interpretable model. **(a)** Input image. The black patch illustrates the small receptive field of **(b)** the modified ResNet-50 backbone. **(c)** The resulting feature map of size $N \times N \times D$, where D is the number of features. **(d)** The class evidence map $\mathbf{A}$ is obtained with C kernels of size 1×1 where C is the number of classes. A sparsity constraint can be placed on $\mathbf{A}$. **(e)** Top to bottom: class probabilities obtained by spatial average pooling and softmax; example class evidence map; suspicious input patches based on the heatmap.

design[1]. Our approach is based on a bag-of-local-features model (BagNet) (Brendel and Bethge, 2019), which has already been shown to be effective in ophthalmology (Ilanchezian et al., 2021). The BagNet relies solely on local evidence, which makes it relevant for classification or detection tasks where the regions of interest are very small and distributed over the image, as is the case in DR. We propose straightforward but effective changes to the BagNet model which lead to fine-grained class evidence maps which directly and correctly reflect the neural network's decision mechanism. Importantly, our approach allows us to enforce sparse heatmaps, which further aids interpretability and allows the model to precisely identify disease related image regions.

2. Developing an interpretable-by-design disease classification network

2.1. Backbone architecture and baseline model

We used a BagNet (Brendel and Bethge, 2019) as a baseline classification model. The BagNet is a variant of ResNet-50 (He et al., 2016) and is obtained by replacing many 3×3 convolutions with 1×1 convolutions and reducing the strides. This leads to a final feature map $\mathbf{F}$ of size $N \times N \times D$ where D is the number of feature channels (Fig. 1c, typically $D = 2048$). Spatial average pooling reduces these features to $1 \times D$, and a linear layer then provides the final prediction logits $\mathbf{l}$ of size $1 \times C$ where C is the number of classes.

These architecture modifications in the BagNet have two effects: First, due to the replaced filters, each image pixel has an effective receptive field of size $q \times q$ in the final feature layer. Therefore, the model makes its predictions based on small image patches of size $q \times q$, implicitly. Secondly, reducing the stride in the convolutional layers prevents

1. Our code is available at https://github.com/kdjoumessi/interpretable-sparse-activation

downsampling effects and results in relatively high-resolution (i.e. fine-grained) feature maps compared to the original ResNet.

2.2. Enhancing the architecture with an interpretable decision-making stage

To interpret predictions in the BagNet model described above, one cannot directly examine the final feature maps (Fig. 1c), as these represent high-dimensional features rather than class evidence at each location. Rather, it is necessary to construct activation maps with multiple forward passes for individual image patches of size $q \times q$ since the original BagNet architecture is not fully convolutional.

Therefore, we introduced class evidence layers to obtain actual class evidence maps $\mathbf{A}_c$ per class c with a single forward pass and make the local information representation explicit. To this end, we reorganised two network operations. As the spatial average pooling and the final fully connected layer are sums, they can be swapped without affecting the final logits:

$$\mathbf{l}_c = \sum_{d=1}^{D} w_{dc} \left(\sum_{i,j \leq N} \frac{1}{N^2} \mathbf{F}_d^{ij} \right) = \sum_{i,j \leq N} \frac{1}{N^2} \left(\sum_{d=1}^{D} w_{dc} \mathbf{F}_d^{ij} \right) = \sum_{i,j \leq N} \frac{1}{N^2} \mathbf{A}_c^{ij}. \tag{1}$$

Importantly, Eq. 1 can be implemented by replacing the (swapped) FC layer by a 1×1 convolution with c output channels. The final class-wise evidence maps $\mathbf{A}_c$ directly represent the contribution of individual input patches to the final prediction. The final class score is then obtained by simple spatial averaging (Fig. 1d), resulting in a c-dimensional logits vector. Applying the softmax function finally leads to the class probabilities $\mathbf{p}_y$ (Fig. 1e).

2.3. Introducing sparsity constraints on class evidence maps

We found that the original BagNet produces dense heatmaps with many positive and negative activations, indicating that clinically irrelevant input patches contribute to the prediction. This behavior makes it difficult for a human to discern how the prediction was formed and to efficiently verify its correctness.

By introducing explicit class evidence layers (Sec. 2.2) we can directly place constraints on the class evidence map containing per-patch scores (Fig. 1d) to induce spatial sparsity. To achieve that, we propose to place an ℓ_1 regularisation constraint on the class evidence maps $\mathbf{A}_c$, leading to the following loss function:

$$\text{Loss}(\mathbf{y}, \mathbf{p}_y) = \text{CE}(\mathbf{y}, \mathbf{p}_y) + \lambda \sum_{i,j,c} |\mathbf{A}_c^{ij}|. \tag{2}$$

Here, CE denotes the cross-entropy and $\mathbf{y}$ are the reference class labels. The sparsity of the activation maps depends on the hyperparameter λ. Enforcing sparsity in class evidence in this way is not a post-hoc measure, but rather forces the classification model to focus on the most relevant image regions. This is particularly suitable for tasks such as DR grading where the detection and characterisation of few lesions in the image is sufficient for an accurate diagnostic result and in line with clinical workflows.

Table 1: Classification performance for referable DR detection on the test set.

	Accuracy	**AUC**	**Specificity**	**Sensitivity**
ResNet-50	0.942	0.960	0.993	0.810
Dense BagNet	0.936	0.957	0.991	0.779
Sparse BagNet	0.928	0.937	1.0	0.750

2.4. Advantages of the new architecture and use in a clinical workflow

The proposed modification of the architecture improves the transparency of the model by providing readily interpretable activation maps which show the contribution of each patch to the final prediction without further post-processing. Furthermore, it provides a different class evidence map for each class in a multi-class scenario, directly showing the contribution of each patch to the classification of the input into that class. Importantly, it does so while being less computationally intensive than the original BagNet.

As we will show below, the class activation maps extracted from the sparse BagNet (Fig. 1d) can be upsampled to the input size and overlaid on the input (Fig. 1e) for easy visualisation and interpretation by clinicians. Further, based on activation scores from the class evidence map, suspicious patches (Fig. 1d) can be extracted and presented to the clinician for further investigations (see Sec. 3.4). In contrast to classical saliency maps, one can directly and straightforwardly report how strongly each patch contributes to the network's decision. A clinician can use the global prediction, the class evidence maps, and suspicious patches to either strengthen their trust in the model or reject a decision.

3. Results

3.1. Dataset

We used retinal fundus images from the Kaggle Diabetic Retinopathy challenge (Kaggle, 2015) with reference DR grades ranging from 0 (no DR) to 4 (proliferative DR). We removed poor-quality images from the dataset using an ensemble of EffcientNet models (Tan and Le, 2019) trained on the ISBI2020[2] challenge dataset. The resulting dataset after the quality filtering contained $45,923$ images with class proportions $(0.73, 0.15, 0.08, 0.03, 0.01)$, which we split into training (75%), validation (10%) and test folds (15%). We preprocessed the images by fitting a circular mask to the field of view and cropping its bounding box. All images were resized to 512×512 and the image intensities were normalised by the mean and standard deviation of the training set. For additional analyses, an experienced in-house ophthalmologist provided detailed lesion annotations on a selection of 15 test images.

3.2. Sparse BagNets yield good accuracy on referable DR detection

We first evaluated our method for the clinically relevant case of (binary) referable DR detection (combining class labels $\{0, 1\}$ vs $\{2, 3, 4\}$). We configured the backbone architecture (Sec. 2.1) such that the receptive field size was $q = 33$ as in Ilanchezian et al. (2021). The

2. https://isbi.deepdr.org/challenge2.html

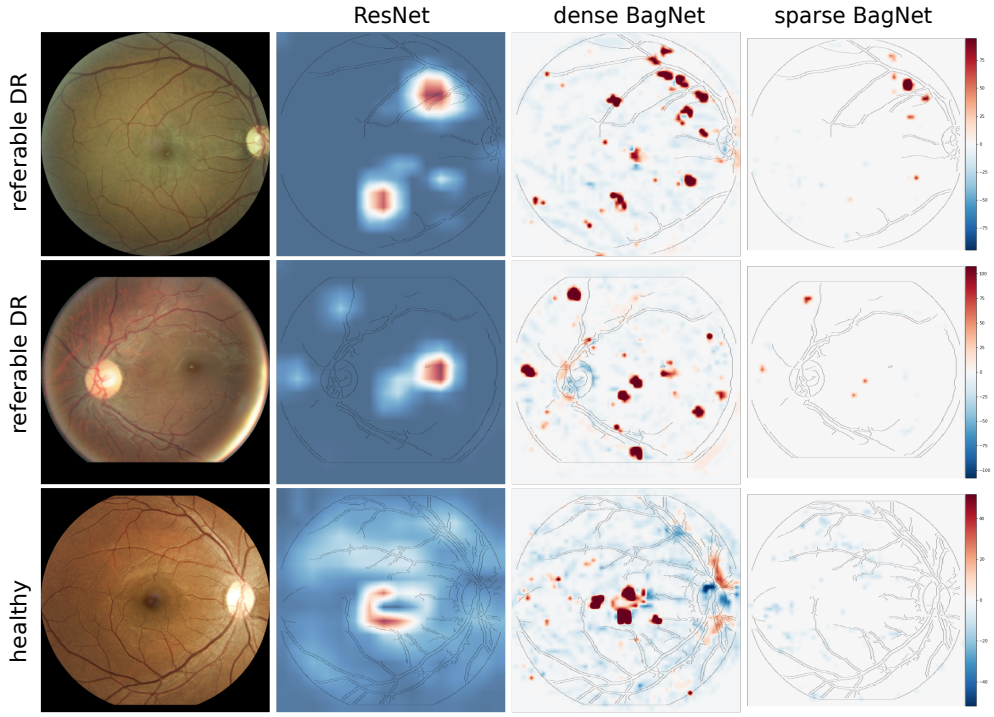

Figure 2: Heatmaps for two example cases with referable DR (top rows) and a healthy case (bottom row). From left to right, heatmaps are shown for the ResNet-50 (using GradCAM), the dense and the sparse BagNet. Red regions provide evidence for the diseased class, while blue regions provide evidence for the healthy class.

regularisation coefficient (Eq. 2) was set to $\lambda = 5{\cdot}10^{-5}$ based on the tradeoff between classification performance on the validation set and sparsity (see App. A). In a realistic application setting, a device manufacturer might define clinically relevant performance thresholds which must be met, and select the maximum sparsity coefficient accordingly.

We compared our sparse BagNet against its dense baseline ($\lambda = 0$) and a ResNet-50 as a black-box state-of-the-art reference. As DR detection has been widely studied classification performance was not our main goal, the training procedures for all models were adopted from Huang et al. (2021), who systematically evaluated hyperparameter choices (see App. B).

We found that the sparse BagNet achieved high accuracy and AUC, which were slightly lower than the respective measures of the dense BagNet and the ResNet50, mainly because of lower sensitivity (Tab. 1). This established that the sparse BagNet was a good candidate for a high-performing interpretable-by-design model, and disease detection performance was not severely hampered by the sparseness penalty on the class activation map.

3.3. Sparsity constraints declutter class evidence maps

We next compared local evidence heatmaps indicating important image features obtained from the different models. For ResNet-50, we used the post-hoc technique GradCAM (Sel-

Table 2: Heatmap evaluation on the test set. The first columns show the local-to-global correspondence for images of healthy and diseased eyes, respectively. The third column shows the localisation precision. For all measures, we report mean (std) per image (higher is better).

	$\mathbf{r^-_{LG}}$	$\mathbf{r^+_{LG}}$	**Precision**
Dense BagNet	0.922 ± 0.03	0.145 ± 0.06	0.219 ± 0.1
Sparse BagNet	0.991 ± 0.04	0.374 ± 0.33	0.791 ± 0.1

varaju et al., 2020). For the BagNet variants, we used class evidence maps directly from the penultimate layer (Sec. 2.2). The ResNet heatmaps were coarse due to the model's large receptive field and some highlighted regions were similar to regions identified by the dense BagNet (Fig. 2, more examples in App. C). However, as the ResNet's heatmaps do not represent the specific local contribution to the model's decision-making process, we focused on the inherently interpretable BagNet versions for further analysis.

Interestingly, the constraints we imposed on our model led to much sparser heatmaps compared to the original BagNet (see right columns in Fig. 2), showing that the decision was formed from few small regions of the retinal fundus. These regions seemed to be mostly a subset of the salient regions used by the dense model. On healthy images, the sparse model led to an almost complete absence of positive activations, in contrast to the mix of positive and negative evidence suggested by the dense model (see bottom row in Fig. 2).

To assess this quantitatively, we measured how consistently the local class evidence (i.e. the heatmap values) corresponded to the global model prediction. For healthy images, we counted all pixels with negative scores and calculated the ratio of pixels with negative scores among all pixels with non-zero scores, which we call local-to-global correspondence r^-_{LG}. This confirmed the qualitative assessment (dense vs. sparse BagNet: 0.922 vs. 0.991; Tab. 2). The same analysis for diseased eyes also further showed a large increase in local-to-global correspondence r^+_{LG} compared to the dense model (0.145 vs. 0.374; Tab. 2). However, on diseased eyes, there remained a large proportion of evidence for healthy tissue, likely because much of the fundus background did not contain any lesions.

3.4. High evidence regions correspond to lesions in sparse BagNet

To assess whether the highlighted regions were clinically relevant for diagnosing DR, we quantified the precision of the BagNets' heatmaps at localising DR lesions on the subset of 15 clinically annotated test images. On these, we extracted input patches with positive scores and calculated precision as the proportion of patches that contained a lesion.

The dense BagNet model contained many positive activations in healthy areas without lesions, resulting in low lesion localisation precision (0.219; Tab. 2). In contrast, the sparse BagNet showed considerably increased precision, almost exclusively extracting patches with lesions (0.791). When we visually inspected the patches identified by the sparse model on the annotated images (Fig. 3), we found that almost all patches with positive scores (red boxes, magnified on the right) contained suspicious spots.

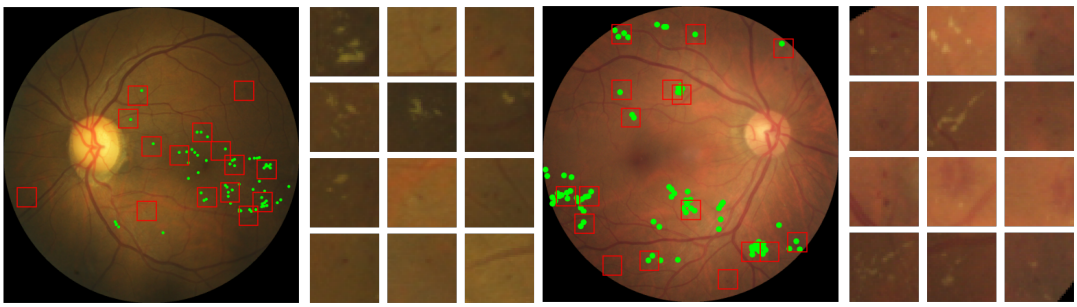

Figure 3: Lesions extracted from the class activation map from the sparse BagNet on two referable DR examples. The green markers indicate reference lesion annotations, whereas the red boxes denote suspicious patches identified by the sparse model (enlarged on the right sorted with decreasing evidence scores).

In fact, when we showed the few seemingly false positive patches (that did not contain an annotation in form of a green marker) to the clinician a second time, she determined that almost all of them likely contain a lesion that had been missed in the original clinical screening.

3.5. Sparse BagNets enhance interpretability for multi-class DR grading

Finally, we applied our method to the multi-class setting of DR grading, where individual severity grades were predicted. We used the same training parameters as for the binary task and set the number of output classes to 5. We set the regularisation coefficient of the sparse model to $\lambda = 6 \cdot 10^{-6}$, again choosing an appropriate accuracy trade-off (see App. A). We found that the dense and sparse models achieved comparable accuracy (resp. 0.864, and 0.850) to the baseline ResNet50 model (0.862).

Again, our sparse BagNet model led to more focused heatmaps that were generally consistent with the predicted class (Fig. 4, more examples in App. D). For the example shown in Fig. 4, a clinician retrospectively confirmed that the image appeared healthy except for diffuse bleeding in line with moderate DR in the area highlighted by our sparse model.

Interestingly, further analysis also helped us to uncover a failure mode of the sparse BagNet: We noticed that sparse BagNets always failed to detect grade 3 and most often grade 4 DR cases (Fig. 4). Instead, it tended to classify these cases as moderate DR (grade 2), likely because the sparse BagNet architecture was not designed to detect the larger lesions occurring in these grades.

4. Discussion and Conclusion

In this paper, we proposed an inherently interpretable classification model which provides sparse high-resolution class evidence maps. Enforcing sparse activations directly caused fewer input regions to contribute to the classifier decision. We showed that the remaining relevant regions in the sparse model identified lesions with high precision, which is a considerable advance of classical saliency map techniques (Saporta et al., 2022). Further,

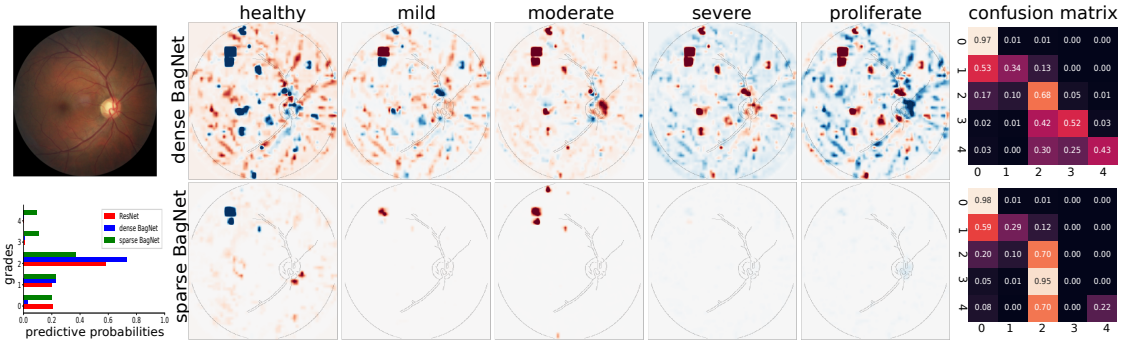

Figure 4: Application to multi-class DR detection shows usefulness of class-specific sparse activation maps of the sparse model over the dense model (bottom row vs. top row; middle column). The example image with moderate DR and predicted probabilities are shown on the left. The confusion matrices (right) show that for the sparse model, severe DR is systematically graded as moderate DR.

healthy images yielded heatmaps with consistently negative evidence. The sparse model was therefore easier and potentially less time-consuming to inspect.

Preliminary feedback from our clinical collaborators suggests that our approach can be a useful tool to verify predictions, understand failure modes of and facilitate their trust in the ML model. Interestingly, they also found the predicted bounding boxes helpful for guiding their attention to subtle anomalies otherwise missed. This suggests future research on ideal ways to let clinicians interact with our model in a human-in-the-loop setting. In a next step, we also plan to apply our approach to other problem settings with local regions of interest such as breast or lung cancer screening.

Acknowledgments

This work was supported by the German Science Foundation (BE5601/8-1 and the Excellence Cluster 2064 “Machine Learning — New Perspectives for Science”, project number 390727645), the Carl Zeiss Foundation in the project “Certification and Foundations of Safe Machine Learning Systems in Healthcare” and the Hertie Foundation. The authors thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting Kerol Djoumessi and Indu Ilanchezian. We also thank Sarah Müller for help with image preprocessing, and Murat Seçkin Ayhan for providing an automated quality grading tool.

References

Wejdan L. Alyoubi, Wafaa M. Shalash, and Maysoon F. Abulkhair. Diabetic retinopathy detection through deep learning techniques: A review. *Informatics in Medicine Unlocked*, 20, 2020. ISSN 2352-9148.

Nishanth Arun, Nathan Gaw, Praveer Singh, Ken Chang, Mehak Aggarwal, Bryan Chen, Katharina Hoebel, Sharut Gupta, Jay Patel, Mishka Gidwani, Julius Adebayo, Matthew D. Li, and Jayashree Kalpathy-Cramer. Assessing the Trustworthiness of Saliency Maps for Localizing Abnormalities in Medical Imaging. *Radiology: Artificial Intelligence*, 3(6), 2021.

Valentyn Boreiko, Indu Ilanchezian, Murat Seçkin Ayhan, Sarah Müller, Lisa M Koch, Hanna Faber, Philipp Berens, and Matthias Hein. Visual explanations for the detection of diabetic retinopathy from retinal fundus images. In *Proc. Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pages 539–549, 2022.

Wieland Brendel and Matthias Bethge. Approximating CNNs with Bag-of-local-Features models works surprisingly well on ImageNet. In *Proc. International Conference on Learning Representations (ICLR)*, 2019.

Mohamed Chetoui and Moulay A. Akhloufi. Explainable Diabetic Retinopathy using EfficientNET. pages 1966–1969. IEEE, July 2020. doi: 10.1109/EMBC44109.2020.9175664. ISSN: 2694-0604.

Jeffrey De Fauw, Joseph R. Ledsam, Bernardino Romera-Paredes, Stanislav Nikolov, Nenad Tomasev, Sam Blackwell, Harry Askham, Xavier Glorot, Brendan O'Donoghue, Daniel Visentin, van den Driessche, et al. Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nature Medicine*, 24(9):1342–1350, 2018.

Cristina González-Gonzalo, Bart Liefers, Bram van Ginneken, and Clara I. Sánchez. Iterative augmentation of visual evidence for weakly-supervised lesion localization in deep interpretability frameworks: Application to color fundus images. *IEEE Transactions on Medical Imaging*, 39(11):3499–3511, 2020.

Thomas Grote and Philipp Berens. On the ethics of algorithmic decision-making in healthcare. *Journal of medical ethics*, 46(3):205–211, 2020.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

Yijin Huang, Li Lin, Pujin Cheng, Junyan Lyu, Roger Tam, and Xiaoying Tang. Identifying the key components in resnet-50 for diabetic retinopathy grading from fundus images: a systematic investigation. *arXiv preprint arXiv:2110.14160*, 2021.

ICO. International council of ophthalmology (ico) guidelines for diabetic eye care, 2017. URL https://icoph.org/eye-care-delivery/diabetic-eye-care/.

Indu Ilanchezian, Dmitry Kobak, Hanna Faber, Focke Ziemssen, Philipp Berens, and Murat Seçkin Ayhan. Interpretable gender classification from retinal fundus images using bagnets. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 477–487. Springer, 2021.

Kaggle. Kaggle competition on diabetic retinopathy detection, 2015. URL https://www.kaggle.com/c/diabetic-retinopathy-detection/data. Accessed: 2022-11-30.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017.

Ella Mahoro and Moulay A. Akhloufi. Applying Deep Learning for Breast Cancer Detection in Radiology. *Current Oncology (Toronto, Ont.)*, 29(11):8767–8793, 2022.

Yu E Nesterov. A method for solving the convex programming problem with convergence rate $\mathcal{O}\left(\frac{1}{k^2}\right)$. In *Dokl. Akad. Nauk SSSR,*, volume 269, pages 543–547, 1983.

Alexandros Papadopoulos, Fotis Topouzis, and Anastasios Delopoulos. An interpretable multiple-instance approach for the detection of referable diabetic retinopathy in fundus images. *Scientific Reports*, 11(1):14326, 2021.

Mihir Rao, Michelle Zhu, and Tianyang Wang. Conversion and implementation of state-of-the-art deep learning algorithms for the classification of diabetic retinopathy. *ArXiv*, abs/2010.11692, 2020.

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.

Adriel Saporta, Xiaotong Gui, Ashwin Agrawal, Anuj Pareek, Steven QH Truong, Chanh DT Nguyen, Van-Doan Ngo, Jayne Seekins, Francis G Blankenberg, Andrew Y Ng, et al. Benchmarking saliency methods for chest x-ray interpretation. *Nature Machine Intelligence*, 4(10):867–878, 2022.

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. *International Journal of Computer Vision*, 128(2):336–359, 2020.

Shiwen Shen, Simon X Han, Denise R Aberle, Alex A Bui, and William Hsu. An interpretable deep hierarchical semantic convolutional neural network for lung nodule malignancy classification. *Expert Systems with Applications*, 128:84–95, 2019.

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

Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. Striving for simplicity: The all convolutional net. *arXiv preprint arXiv:1412.6806*, 2014.

Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019.

Meysam Tavakoli and Patrick Kelley. *A comprehensive survey on computer-aided diagnostic systems in diabetic retinopathy screening.* 2021.

Qiaoying Teng, Zhe Liu, Yuqing Song, Kai Han, and Yang Lu. A survey on the interpretability of deep learning in medical diagnosis. *Multimedia Systems*, 28(6):2335–2355, 2022.

Daniel Shu Wei Ting, Gemmy Chui Ming Cheung, and Tien Yin Wong. Diabetic retinopathy: global prevalence, major risk factors, screening practices and public health challenges: a review. *Clinical & Experimental Ophthalmology*, 44(4), 2016.

Zhiguang Wang and Jianbo Yang. Diabetic Retinopathy Detection via Deep Convolutional Networks for Discriminative Localization and Visual Explanation. In *Workshops at the thirty-second AAAI conference on artificial intelligence*. arXiv, 2019.

Tien Y Wong, Jennifer Sun, Ryo Kawasaki, Paisan Ruamviboonsuk, Neeru Gupta, Van Charles Lansingh, Mauricio Maia, Wanjiku Mathenge, Sunil Moreker, Mahi MK Muqit, et al. Guidelines on diabetic eye care: the international council of ophthalmology recommendations for screening, follow-up, referral, and treatment based on resource settings. *Ophthalmology*, 125(10):1608–1622, 2018.

Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016.

Appendix A. Effect of the sparsity hyperparameter λ

The regularisation coefficient λ (see Eq. 2) is the hyperparameter that controls the sparsity of the class-specific activation map in the sparse BagNet. It was chosen based on a tradeoff between performance on the validation set according to each task (Fig. 5). Specifically, we manually selected the highest sparsity coefficient for which the performance did not drop too strongly.

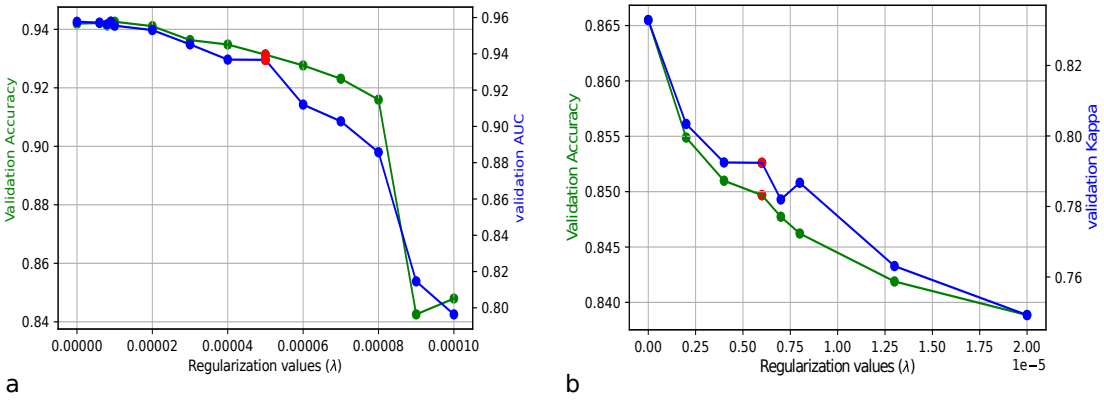

Figure 5: Comparison of validation performance with different regularization values. **(a)** The regularization coefficient λ affects the AUC and accuracy on the binary referable task. **(b)** Same as **(a)**, but for the multiclass task with accuracy and kappa. The red points indicate the selected values, which are a trade-of between sparsity and accuracy.

Appendix B. Training details

We adopted the training regime and hyperparameter choices optimised by Huang et al. (2021), who systematically evaluated relevant hyperparameters for DR detection on fundus images, such as the influence of data preprocessing and augmentation, optimiser and learning rate configurations. The final settings are described below.

We performed data augmentation during training by flipping, rotating, randomly cropping, and translating with a given probability. We also used Krizhevsky color augmentation (Krizhevsky et al., 2017), as suggested by Huang et al. (2021).

To train the networks, we used the cross-entropy loss (unless specified otherwise) as the objective function with the SGD optimizer where the initial learning rate was set to 0.001. Next, the cosine learning schedule was used and the minimum learning rate was set to 0.0001. We also used Nesterov's momentum (Nesterov, 1983) with a constant momentum factor of 0.9 with a weight decay of 0.0005 for regularization.

Models were initialized with weights obtained on the ImageNet and fine-tuned on the Kaggle dataset for 100 epochs with a mini-batch size of 8. The best model was saved on the validation set depending on the task (binary referable DR detection or multiclass DR grading).

For simplicity, and as our goal was not to push the boundaries of classification performance, we did not incorporate features from opposite eyes for image grading (as suggested by Huang et al. (2021)), and also omitted ensembling multiple models.

Appendix C. Additional examples of class evidence maps

An additional selection of class evidence maps for correctly classified and misclassified examples is provided in Fig. 6.

Appendix D. Additional results for multiclass setting

The classification performance of our sparse model was comparable to its dense baseline and the ResNet (see Tab. 3). The multiclass task (Fig. 7) shows the advantages of having class-specific activation maps: For example in the last row, small lesions are detected arguing for moderate DR, but larger deteriorations towards the edge of the image argue for proliferate DR. .

	Acc.	**Kappa**
ResNet-50	0.862	0.826
Dense BagNet	0.864	0.830
Sparse BagNet	0.850	0.780

Table 3: Comparison of the classification performances on multiclass DR detection between the proposed approaches and the baseline ResNet-50 model on the test set.

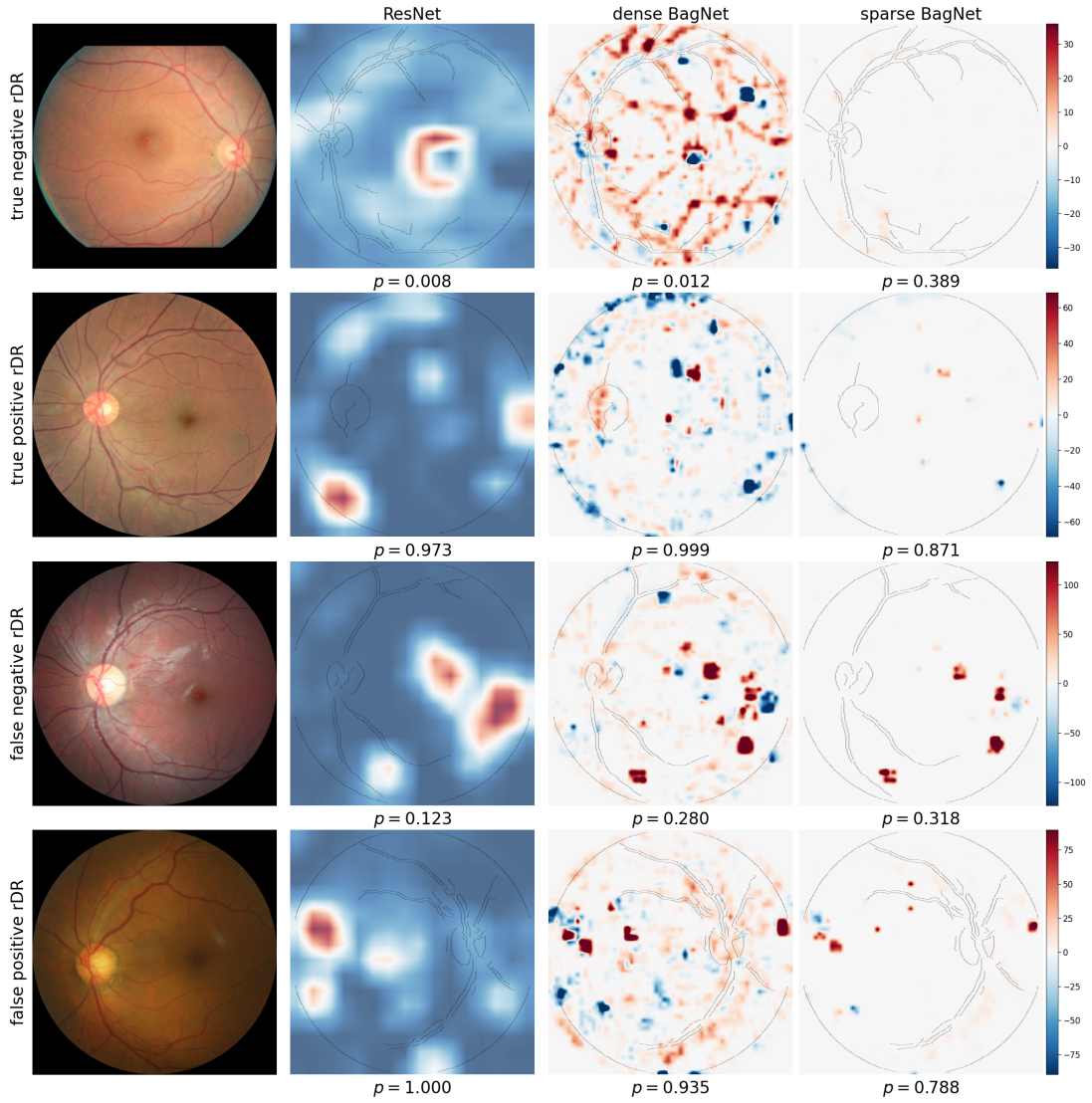

Figure 6: Heatmaps for example of correctly and misclassified cases with referable and healthy DR images. From left to right, heatmaps are shown for ResNet-50 (using GradCAM), the dense and sparse BagNet. Below each heatmap, we also show the predicted probability for referable DR.

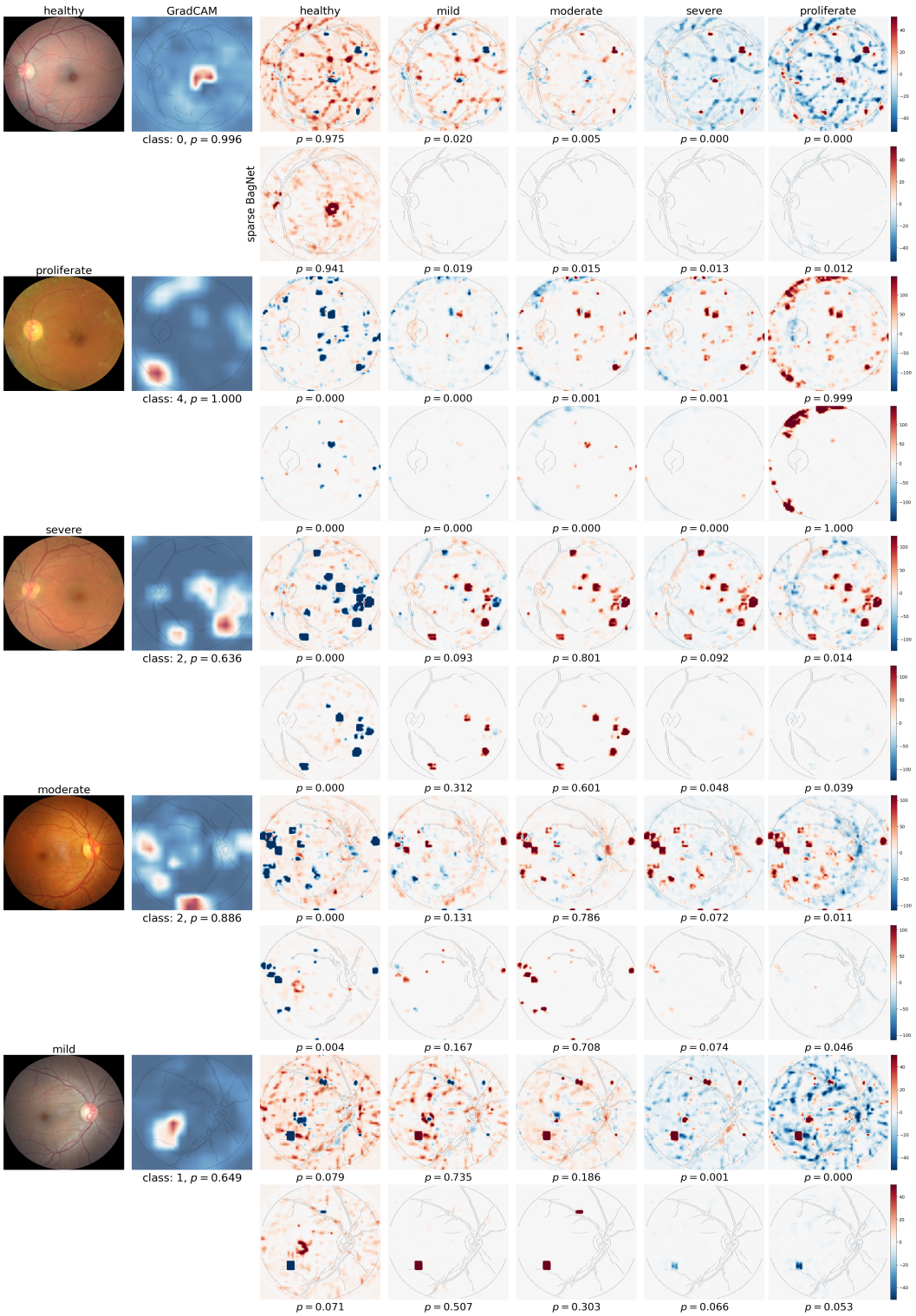

Figure 7: Multi-class evidence maps for images of different DR grades. From left to right, we show the fundus image, the ResNet-50 heatmap (using GradCAM), and the class-specific map for different grades for the dense (top) and sparse (bottom) BagNet. Below each heatmap, we also show the predicted probability for each class, as well as the predicted class and probability for the ResNet-50 model.

PLOS DIGITAL HEALTH

RESEARCH ARTICLE

An inherently interpretable AI model improves screening speed and accuracy for early diabetic retinopathy

Kerol Djoumessi[1,2], **Ziwei Huang**[1,2], **Laura Kühlewein**[3], **Annekatrin Rickmann**[3,4], **Natalia Simon**[5], **Lisa M. Koch**[1,2,6], **Philipp Berens**[1,2*‡]

1 Hertie Institute for AI in Brain Health, University of Tübingen, Tübingen, Germany, **2** Tübingen AI Center, University of Tübingen, Tübingen, Germany, **3** University Eye Hospital, University of Tübingen, Tübingen, Germany, **4** Eye Clinic Sulzbach, Knappschaft Hospital Saar, Sulzbach, Germany, **5** Black Forest Eye Clinic, Endingen, Germany, **6** Department of Diabetes, Endocrinology, Nutritional Medicine and Metabolism UDEM, Inselspital, Bern University Hospital, University of Bern, Bern, Switzerland

‡ Current address: Hertie Institute for AI in Brain Health, University of Tübingen, Tübingen, Germany
* philipp.berens@uni-tuebingen.de

Citation: Djoumessi K, Huang Z, Kühlewein L, Rickmann A, Koch LM, Berens P (2025) An inherently interpretable AI model improves screening speed and accuracy for early diabetic retinopathy. PLOS Digit Health 4(5): e0000831. https://doi.org/10.1371/journal.pdig.0000831

Editor: Po-Chih Kuo, National Tsing-Hua University: National Tsing Hua University, TAIWAN

Received: January 07, 2025

Accepted: March 19, 2025

Published: May 12, 2025

Peer Review History: PLOS recognizes the benefits of transparency in the peer review process; therefore, we enable the publication of all of the content of peer review and author responses alongside final, published articles. The editorial history of this article is available here: https://doi.org/10.1371/journal.pdig.0000831

Data availability statement: The implementation of our sparse BagNet model is

Abstract

Diabetic retinopathy (DR) is a frequent complication of diabetes, affecting millions worldwide. Screening for this disease based on fundus images has been one of the first successful use cases for modern artificial intelligence in medicine. However, current state-of-the-art systems typically use black-box models to make referral decisions, requiring post-hoc methods for AI-human interaction and clinical decision support. We developed and evaluated an inherently interpretable deep learning model, which explicitly models the local evidence of DR as part of its network architecture, for clinical decision support in early DR screening. We trained the network on 34,350 high-quality fundus images from a publicly available dataset and validated its performance on a large range of ten external datasets. The inherently interpretable model was compared to post-hoc explainability techniques applied to a standard DNN architecture. For comparison, we obtained detailed lesion annotations from ophthalmologists on 65 images to study if the class evidence maps highlight clinically relevant information. We tested the clinical usefulness of our model in a retrospective reader study, where we compared screening for DR without AI support to screening with AI support with and without AI explanations. The inherently interpretable deep learning model obtained an accuracy of .906 [.900–.913] (95%-confidence interval) and an AUC of .904 [.894–.913] on the internal test set and similar performance on external datasets, comparable to the standard DNN. High evidence regions directly extracted from the model contained clinically relevant lesions such as microaneurysms or hemorrhages with a high precision of .960 [.941–.976], surpassing post-hoc techniques applied to a standard DNN. Decision support by the model highlighting high-evidence regions in the image improved screening accuracy for difficult decisions and improved screening speed. This shows that inherently interpretable deep learning models can provide clinical decision support while obtaining state-of-the-art performance improving human-AI collaboration.

available at GitHub (https://github.com/kdjoumessi/Sparse-BagNet_clinical-validation). The annotations performed for this study on selected Kaggle database images, the study data, and the analysis are available in the same GitHub repository.

Funding: This work was supported by a grant of the Hertie Foundation to PB, grants from the German Research Foundation (BE5601/8-1 to PB; Excellence Cluster 2064 "Machine Learning — New Perspectives for Science", project number 390727645 to PB), a grant from the Carl Zeiss Foundation ("Certification and Foundations of Safe Machine Learning Systems in Healthcare" to LK). Furthermore, the International Max Planck Research School for Intelligent Systems (IMPRS-IS) supported KD. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Author summary

AI systems designed to support clinical decision making use black-box deep learning models for many medical applications. This includes AI-based screening systems for diabetic retinopathy, a sight threatening complication of diabetes. This hinders clinical uptake of such methods, as clinicians and patients do not have a way to validate the AI systems decisions. Sometimes, post-hoc methods are used to generate heatmaps that supposedly explain the AI systems decision. However, these methods are problematic as the generated explanations do not reflect the actual decision-making process of the model and are prone to spurious correlations. In our paper, we take a big step forward for enabling trustworthy AI systems for supporting clinical decision making in screening for diabetic retinopathy: We introduce an inherently interpretable deep learning model which provides human-understandable explanations for its decisions. The model combines the power of deep learning with the interpretability of simpler models such as logistic regression by computing an explicit evidence map. This map forms the basis of the model's decisions, alleviating the issues of post-hoc techniques. We validate the clinical potential of this model for improving diabetic retinopathy screening showing highlighting regions with high disease evidence during clinical grading decreased the grading time significantly and improved grading accuracy for difficult borderline cases.

Introduction

Diabetic retinopathy (DR) screening has been one of the first successful use cases for artificial intelligence (AI) in medicine [1], promising fast, cost-effective access even where insufficient clinical personnel is available. By now, multiple AI systems have received regulatory clearance [2,3] and have been found useful to triage patients not requiring specialist attention and those with vision-threatening DR, potentially contributing to increased screening adherence [4].

However, current state-of-the-art models use black-box deep learning approaches to make referral decisions, providing clinicians only with limited binary recommendations to either refer a patient for further examination or not. Yet, the performance of current systems still typically makes some level of human grader verification necessary [3], which could be guided by an useful explanation of the AI system's decision. Also, clinical implementation would benefit from clinicians being able to understand the rationale behind the recommendation of the algorithm [5–7].

Typically, an AI system's decision are explained with heatmaps obtained post-hoc using gradient-based approaches [8–10]. However, such explanations are not trustworthy, as the produced heatmaps do not reflect the actual decision-making process of the model, and are prone to spurious correlations [11]. Therefore, their results cannot be easily integrated into the clinical decision-making process [7,12].

We address this issue and validate an inherently interpretable deep learning architecture for providing clinical decision support for screening for early DR in a retrospective reader study. Our approach uses a deep learning architecture called sparse BagNets [13,14], which explicitly models the local evidence for the presence of DR as part of its network architecture (Fig 1B). Most studies so far have considered the task of screening for moderate non-proliferative DR or more advanced stages [1], although even mild non-proliferative diabetic retinopathy (NPDR) is recommended for close monitoring and careful control of hyperglycemia [15,16]. We reasoned that the benefit of AI-based explanations and decision support would be most clearly visible for this challenging diagnostic task. Trained on a large

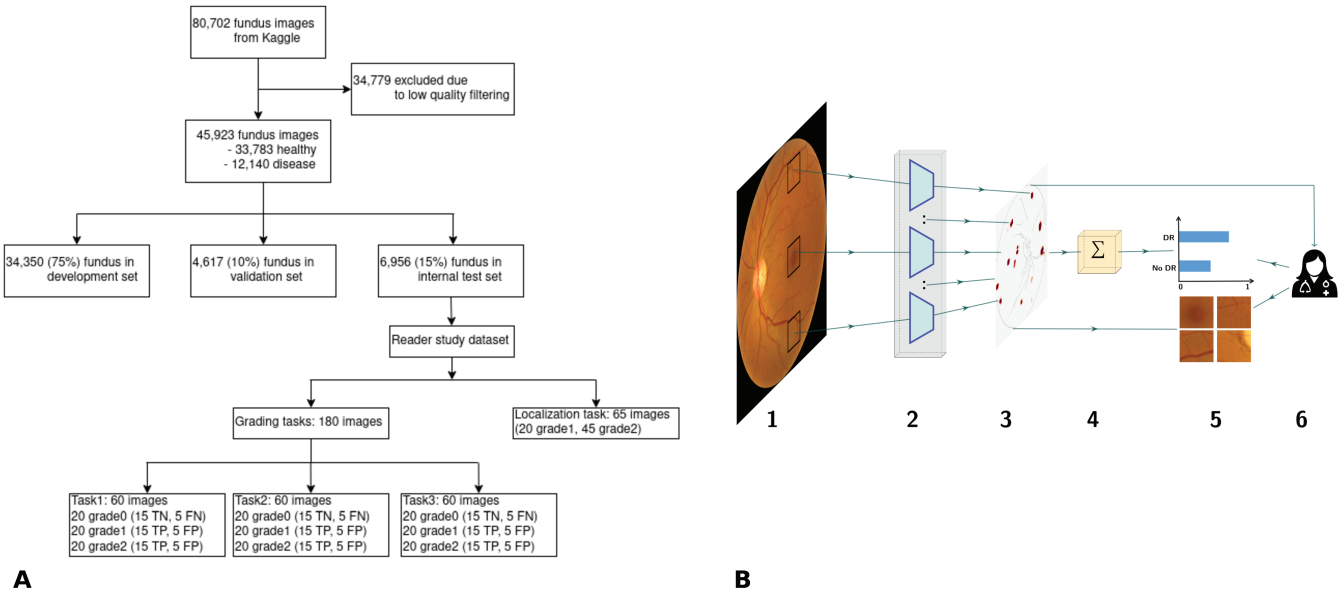

Fig 1. Overview of the development data and proposed inherently interpretable deep learning framework evaluated in this study. (A) Summary of the development dataset used to build the model, as well as the data used in the retrospective reader study. (B) Sparse BagNet architecture. 1. As a preliminary step, the retinal fundus image is implicitly split into many overlapping small patches of size 33 × 33. 2. All patches are fed to the model backbone, which processes them in parallel. 3. The BagNet backbone generates a heatmap that depicts the local disease evidence of individual patches. 4. The values of the heatmap are averaged and used as the final logit for classification. 5./6. The logits are fed into a softmax function which provides the probability distribution of the output, and then patches of suspect regions based on the heatmaps can be requested and viewed by a clinician to understand the classification results.

https://doi.org/10.1371/journal.pdig.0000831.g001

publicly available dataset, our model shows high specificity and sufficient sensitivity in detecting mild DR across a large array of datasets. Importantly, we show that the obtained class evidence maps highlight clinically relevant lesions such as microaneurysms or hemorrhages with high precision, making them useful for verifying the AI system's decisions. Finally, we show that the system can be effectively used to guide clinical decision-making, leading to 17.5% improvement in diagnostic accuracy for mild DR and overall about ≈ 25% improvement in screening time.

Methods

Dataset description and data preparation

We used eleven publicly available retinal image datasets, consisting of color fundus images from various sources, to develop and evaluate an inherently interpretable deep learning model for early DR detection (Table 1). For all datasets, fundus images had assigned reference grades based on the International Clinical Diabetic Retinopathy classification scale [17], which provides a grading scheme ranging from 0 (no DR), 1 (mild NPDR), 2 (moderate NPDR), 3 (severe NPDR) to 4 (proliferative DR) according to DR severity. As our goal was to develop an AI system for early DR screening, we combined class level {0} vs {1,2,3,4}. At stage 1, DR is in most cases asymptomatic, and challenging to detect even for experienced ophthalmologists. As all fundus datasets were fully anonymous, no approval from an Ethics Board was needed for this part of the study.

Table 1. Summary of the internal and external validation datasets used to evaluate the models. "Origin" refers to the country where the data was collected. "Lesion" refers to the number of images in the dataset with lesion annotations. The Kaggle dataset (first row, shaded in gray) is the internal dataset used to evaluate the model, while the other datasets were used for external validation to assess the generalization properties of the trained model.

Dataset	Origin	Number of images			Lesion
		All	Healthy	DR	
Kaggle [18]	USA	6,956	5,118	1,838	65
IDRiD [19]	India	512	168	348	81
E-Ophtha [20]	France	434	260	174	174
FGA-DR [21]	UAE	1,841	101	1,740	1,740
DIARETDB1 [22]	Finland	89	05	84	84
DDR [23]	China	12,513	6,265	6,248	755
DR2 [24]	Brazil	445	300	145	-
APTOS [25]	India	3,662	1,805	1,857	-
FCM-UNA [26]	Paraguay	757	187	570	-
Messidor-1 [27]	France	1,200	546	654	-
Messidor-2 [27,28]	France	1,744	1,017	727	-

https://doi.org/10.1371/journal.pdig.0000831.t001

Development dataset.

The dataset used to develop the inherently interpretable deep learning model was obtained from the Kaggle Diabetic Retinopathy challenge [18] which initially contained records of 44,351 subjects with 88,702 retinal fundus images from both eyes (Fig 1A). This dataset was originally provided by EyePacs Inc., a diabetes screening program in California. A comparable dataset also obtained from EyePacs Inc. included ethnicity information and contained about 70% images from patients with Latin American ethnicity [29]. We automatically quality filtered the fundus images using an ensemble of 10 EfficientNets models [30] trained on the DeepDRiD dataset [31]. This model achieved a quality filtering accuracy of 87.5% [32]. After quality filtering, we retained 45,923 images from 28,984 subjects for training, with 73% of images in the healthy class and 27% in the DR class. The dataset was split into training, validation, and test folds with 75%, 10%, and 15% of images, respectively, making sure that all images from the same subject were allocated to the same fold. The training fold was used for model fitting, the validation fold for model selection and hyperparameter tuning, and the test fold for internal evaluation.

To evaluate the explanations provided by the explainable sparse BagNet model, three ophthalmologists (authors AR, LaK, and NS with 5, 9, and 14 years of experience respectively) marked the location of DR-related lesions on 65 randomly selected fundus images from the test set (20 grade 1 and 45 grade 2) using a custom-written annotation browser interface (S1 Fig) based on the Python web framework Django, version 4.2.1, with a secure PostgreSQL database, version 15.3, and a Javascript front-end (available at https://github.com/berenslab/retimgtools/releases/tag/v1.1.0). Annotators were asked to mark "Microaneurysms (MA)", "Hemorrhages (HE)", "Exudates (EX)", "Soft Exudates (SE)" or "Other" for lesions visible on the fundus image. We combined the annotations of all graders into a consensus annotation for each image (S1 Table). We also assessed the consistency between ophthalmologists' annotations by calculating the dice between their annotations, showing that annotating DR-related lesions exhaustively is a challenging task (S2 Table).

External datasets.

Additional fundus data sets were obtained from various sources (Table 1) and were used for external evaluation of the model to assess the generalization performance. In addition to

reference DR grades, some of these external datasets [19–23] contained pixel-wise annotations for disease-related lesions. We used these additional annotations to evaluate the performance of the interpretable deep-learning model at localizing DR-related lesions.

Preprocessing.

Raw fundus images were preprocessed by cropping them to a square size of 512 x 512 pixels using a circle fitting method [33]. Then, image intensities were normalized by the mean and standard deviation of the training set. We applied this preprocessing procedure to all the fundus images from all datasets with the same parameters.

Inherently interpretable deep learning model for Diabetic Retinopathy detection

Architecture.

We trained and evaluated an inherently interpretable deep convolutional neural network (sparse BagNet [13,14]) for early DR detection. The sparse BagNet is an implicitly patch-based model based on bag-of-local features and aggregates local evidence from interpretable heatmaps to make predictions (Fig 1B). It takes a two-dimensional fundus image as input (Fig 1B.1) and outputs a binary prediction, which indicates the absence or presence of DR, together with the confidence as the probability score.

In contrast to other deep learning models, the sparse BagNet architecture is designed to be inherently interpretable, as the input image is implicitly split into many small, overlapping patches (size q = 33x33 pixels corresponding to the size of the model's effective receptive field with stride s = 8; Fig 1B.1), which are independently processed in parallel (Fig 1B.2) to compute the local evidence for the presence of DR. The patchwise predicted local evidence values are combined into a single class evidence map corresponding to a downsampled version of the input image (Fig 1B.3), which then is aggregated using average pooling and passed through a softmax function (Fig 1B.4) to output the probability distribution of DR (Fig 1B.5). Crucially, we employ a ℓ_1-penalty on the local evidence to encourage a sparse class evidence map.

After inference, the model can support screening not only with the final prediction but also with the class evidence map (Fig 1B.3) highlighting the contribution of small local regions to the final prediction. To this end, the evidence map is upsampled to the full image resolution and overlaid on the input image. In contrast to post-hoc gradient-based methods [11], the class evidence map provided by the sparse BagNet is a transparent part of the actual decision-making process and faithfully captures the local evidence. We supplement the class evidence map by extracting patches from regions with high DR evidence (Fig 1B.5).

Training procedure.

We trained the model on the training set by minimising the following loss function including the ℓ_1-penalty:

$$L\big((\mathbf{X},\theta),\mathbf{y}\big) = CE\big(f(\mathbf{X},\theta),\mathbf{y}\big) + \lambda \sum_{i,j,c} |\mathbf{A}_c^{ij}|. \tag{1}$$

Here, $\mathbf{X} \in \mathbb{R}^{H\times W\times C}$ denotes the input image with H,W,C being height, width, and the number of channels, CE is the cross-entropy, $\mathbf{y}$ are the reference class labels, f is the model with parameters θ, and $\mathbf{A}_c$ denotes the evidence map of class c. The sparsity of the evidence maps depends on the hyperparameter λ.

We initialized the model with weights pre-trained on ImageNet and then retrained and optimized for accuracy on the Kaggle DR dataset for 100 epochs. We used the stochastic gradient descent optimizer with an initial learning rate of 10^{-3}, and a clipped cosine learning rate scheduler with a minimum value set to 10^{-4}. We performed data augmentation during training by applying random cropping, flipping, color jitter, translation, and rotation following [34]. The sparsity hyperparameter λ was chosen based on the classification accuracy on the validation set (S2 Fig).

Baseline model and post-hoc interpretability

For comparison, we trained a standard black-box ResNet-50 [35] for early onset DR detection using the same training procedure as described above. We evaluated the classical interpretability techniques Integrated Gradients and Guided Backpropagation due to their high performance in identifying clinically validated DR lesions [36].

Clinical user study for AI-based decision support

Study dataset.

The user study was designed to evaluate the usefulness of the explanations provided by the inherently interpretable deep learning model in clinical practice. The dataset for each grading task (see below) consisted of 60 fundus images from the internal test set, where 20 images were sampled from grade 0, grade 1, and grade 2 respectively. For each grade, 15 images were correctly classified by the network and 5 falsely, making this a challenging screening task for clinicians. Thus, the fraction of images with DR in the user study was 66% and the deep learning model achieved an accuracy of 75% by design. Image grading was based solely on the fundus image and AI support, but no additional clinical data were provided.

Study design.

Six trained ophthalmologists with a median clinical experience of 9 years (4–17 years) participated in the reader study (including authors LaK, AR, and NS). We did not perform a formal power calculation. The study consisted of three tasks: In task 1 (referred to as "H"), participants were asked to grade fundus images without AI support (S3 Fig). In task 2 ("H+AI"), participants were additionally provided with the class predicted by the deep learning model and its confidence (S4 Fig). Finally, in task 3 ("H+XAI"), participants were additionally shown model explanations in the form of up to 12 bounding boxes around the regions from the class evidence map with the highest evidence, with bounding boxes matching the effective receptive field size and depicting the local image patches that contribute most to the global class evidence (S5 Fig).

For the three grading tasks, readers were instructed to classify each fundus image into two classes ("No DR" and "DR"). They were told to classify an image as "DR" even if they thought it only contained signs of mild non-proliferative DR (grade 1). None of the readers had access to the true labels. For task 3, readers were told that some bounding box explanations may contain healthy regions, as the algorithm also generated bounding boxes for healthy images erroneously classified as DR by the sparse BagNet model. In addition to the assigned class, we recorded the time it took for the reader to grade each image and asked them to rate their confidence on a scale from 1 to 5. Ethical approval for the study was obtained from the Ethics Committee at the University Hospital Tübingen (Ref No. 249/2023BO2).

A custom-written browser interface based on the Python web framework Django (version 4.2.1) with a secure PostgreSQL database (version 15.3) and a JavaScript front-end

was used to carry out the study (S3 Fig, S4 Fig, S5 Fig). The tool showed the fundus image, and response options and provided a digital magnifier to enlarge small image regions.

Evaluation criteria and statistical analysis

Criteria for evaluating the performance of the inherently interpretable deep learning model were specified before the start of the study based on previous work [13]. We evaluated three aspects of the model's quality:

1. DR screening performance compared to a regular deep learning model, within and across datasets.
2. The quality of the class evidence maps and derived bounding boxes in terms of lesion localization.
3. The usefulness of the inherently interpretable deep-learning model and the derived bounding boxes for decision support.

DR screening performance.

The primary measure of DR screening performance was the accuracy of the model for early DR detection using the reference labels. Additionally, we evaluated the area under the receiver-operating curve (AUC), sensitivity, specificity, and precision. All measures were computed on the internal test set as well as on the ten external datasets (Table 1). The model was not retrained or fine-tuned before assessment on the external datasets. All measures were computed using the scikit-learn package (v 1.0.2) and confidence intervals were computed using a bootstrap procedure with 1000 unstratified resamples [37].

Quality of class evidence maps.

To measure the quality of the class evidence maps and the derived bounding boxes for lesion localization, we calculated the proportion of highlighted regions (regions within the bounding box) that contained annotated lesions ("localization precision"). To this end, we used the annotations collected for this study on 65 images from the test set, as well as those external datasets containing pixel-level annotations (Table 1). We did not evaluate the fraction of lesions detected by our model ("recall"), as we did not train the model for lesion detection, and diagnostic support does not require an exhaustive detection of all lesions.

Statistical analysis of decision support.

We measured the performance of the readers in our clinical user study as the accuracy of the reader's decision with respect to the reference labels. To assess the effect of the task and DR reference grade statistically, we fit the responses with a generalized linear model (R, function *glm*, v 4.0.3) with predictor *task* or with predictors *task* and *DR grade* including interactions. If we found significant predictors at the $\alpha = 0.05$ level, we computed the marginal means and 95%-confidence intervals (package *emmeans*, v 1.5.3) as well as the respective contrasts between conditions for post-hoc testing. Tukey's method was used for correcting for multiple comparisons. We used the same procedure for analyzing the measured grading time and the reported confidence, but used a linear model (function *lm*) instead.

Role of the funding source

The funders of this work had no role in the study design, collection, analysis, and interpretation of data, the writing of the report, nor in the decision to submit the paper for publication.

Results

We trained and evaluated an inherently interpretable deep learning model ("sparse BagNet") for early DR screening (Fig 1B). We first evaluated screening performance for early DR against the state-of-the-art non-interpretable black-box model ("ResNet50") on the internal test set of the development dataset and on a large number of additional datasets (see Table 2). The sparse BagNet performed well and was comparable to the state-of-the-art model on the internal test set (accuracy: 0.906, 95% CI [0.900–0.913]; AUC: 0.904 [0.894–0.913]; sensitivity: 0.709 [0.688–0.729]; specificity: 0.977 [0.973–0.981]; precision: 0.918 [0.903–0.932]) and generalized well to a number of external datasets (Table 2).

The key advantage of our inherently interpretable model is that the local disease evidence is explicitly represented in a class evidence map (Fig 1B.3 and Fig 2B). During training, the class evidence map is encouraged to be sparse, such that the final loss function balances prediction accuracy and an interpretable map. For the model studied above, the regularization parameter trading-off accuracy and sparseness was heuristically chosen such that sparseness was encouraged at a minimal loss of accuracy (S2 Fig). At each location in the class activation map, the color indicates the model output for an individual image patch. We detected the regions with the highest evidence and placed bounding boxes corresponding to the patch size around these points (Fig 2A).

Table 2. Summary of the classification performance with confidence intervals (CIs) computed at 95% using bootstrapping (n=1000). "AUC" refer to the receiver-operating curve. "Loc Bag" and "Loc GBP" respectively refer to the localization precision of the sparse BagNet and Guided Backpropagation on ResNet-50 at localizing lesions from annotated images. For each dataset, the first row shows the performance of the interpretable sparse BagNet model, while the second row shows the performance of the baseline black-box ResNet-50 model. The Kaggle dataset (first row) is the internal dataset used to train and evaluate the model, while the other datasets were used for external validation to assess the generalization properties of the trained model. The low classification performance on the FCM-UNA and FGA-DR datasets can be explained by the relatively low quality of most images in the FCM-UNA dataset and the large intensity variation of the FGA-DR dataset (S6 Fig). The low localization precision (0.664) on the E-Ophtha dataset is likely due to annotations only being provided for "Microaneurysms" and "Exudate" lesions, while the images could contain other DR-related lesions.

Dataset		Accuracy	AUC	Sensitivity	Specificity	Precision	Loc_{Bag}	Loc_{GBP}
Kaggle	Bag.	.906 (.900 - .913)	.904 (.894 - .913)	.709 (.688 - .729)	.977 (.973 - .981)	.918 (.903 - .932)	.941	-
	Res.	.914 (.907 - .921)	.935 (.927 - .943)	.765 (.745 - .784)	.967 (.962 - .972)	.894 (.878 - .908)	-	.656
		.891 (.864 - .917)	.879 (.838 - .913)	.951 (.927 - .972)	.768 (.699 - .828)	.895 (.861 - .925)	.804	-
IDRiD		.882 (.851 - .909)	.864 (.822 - .902)	.963 (.942 - .981)	.714 (.639 - .781)	.875 (.84 - .908)	-	.140
		.903 (.864 - .917)	.944 (.838 - .913)	.920 (.927 - .972)	.892 (.699 - .828)	.851 (.861 - .925)	.656	-
E-Ophtha		.933 (.851 - .909)	.972 (.822 - .902)	.966 (.942 - .981)	.912 (.639 - .781)	.880 (.840 - .908)	-	.030
		.799 (.781 - .819)	.789 (.752 - .823)	.811 (.793 - .830)	.594 (.500 - .687)	.972 (.963 - .980)	.872	-
FGA-DR		.763 (.743 - .781)	.816 (.768 - .858)	.764 (.743 - .783)	.743 (.653 - .819)	.981 (.973 - .987)	-	.336
		.831 (.753 - .899)	.931 (.870 - .981)	.821 (.733 - .898)	1	1	.881	-
DIARETDB1		.742 (.652 - .831)	.811 (.715 - .900)	.738 (.640 - .829)	.800 (.333 - 1.00)	.984 (.950 - 1.00)	-	.000
		.825 (.818 - .832)	.926 (.922 - .931)	.669 (.657 - .681)	.980 (.977 - .984	.971 (.966 - .976)	.965	-
DDR		.887 (.881 - .892)	.963 (.960 - .966)	.800 (.790 - .810)	.973 (.968 - .977)	.967 (.962 - .972)	-	.249
		.879 (.847 - .908)	.922 (.889 - .951)	.662 (.584 - .742)	.983 (.968 - .997)	.950 (.905 - .990)	-	
DR2		.876 (.845 - .906)	.866 (.825 - .905)	.669 (.591 - .742)	.977 (.959 - .993)	.933 (.884 - .975)	-	
		.973 (.968 - .979)	.995 (.992 - .996)	.982 (.975 - .987)	.965 (.956 - .973)	.966 (.958 - .974)	-	
APTOS		.949 (.942 - .956)	.972 (.965 - .978)	.942 (.931 - .952)	.956 (.946 - .965)	.956 (.947 - .966)	-	
		.773 (.744 - .802)	.936 (.918 - .952)	.702 (.664 - .738)	.989 (.972 - 1.00)	.995 (.987 - 1.00)	-	
FCM-UNA		.877 (.853 - .900)	.967 (.954 - .979)	.840 (.811 - .868)	.989(.971 - 1.00)	.996 (.989 - 1.00)	-	
		.889 (.871 - .907)	.943 (.929 - .955)	.832 (.804 - .859)	.958 (.939 - .974)	.959 (.941 - .975)	-	
Messidor-1		.893 (.876 - .909)	.954 (.942 - .965)	.852 (.823 - .878)	.943 (.923 - .963)	.947 (.928 - .964)	-	
		.829 (.812 - .847)	.876 (.859 - .894)	.750 (.719 - .785)	.886 (.865 - .906)	.825 (.794 - .853)	-	
Messidor-2		.851 (.835 - .869)	.925 (.912 - .938)	.794 (.763 - .823)	.893 (.875 - .913)	.841 (.815 - .868)	-	

https://doi.org/10.1371/journal.pdig.0000831.t002

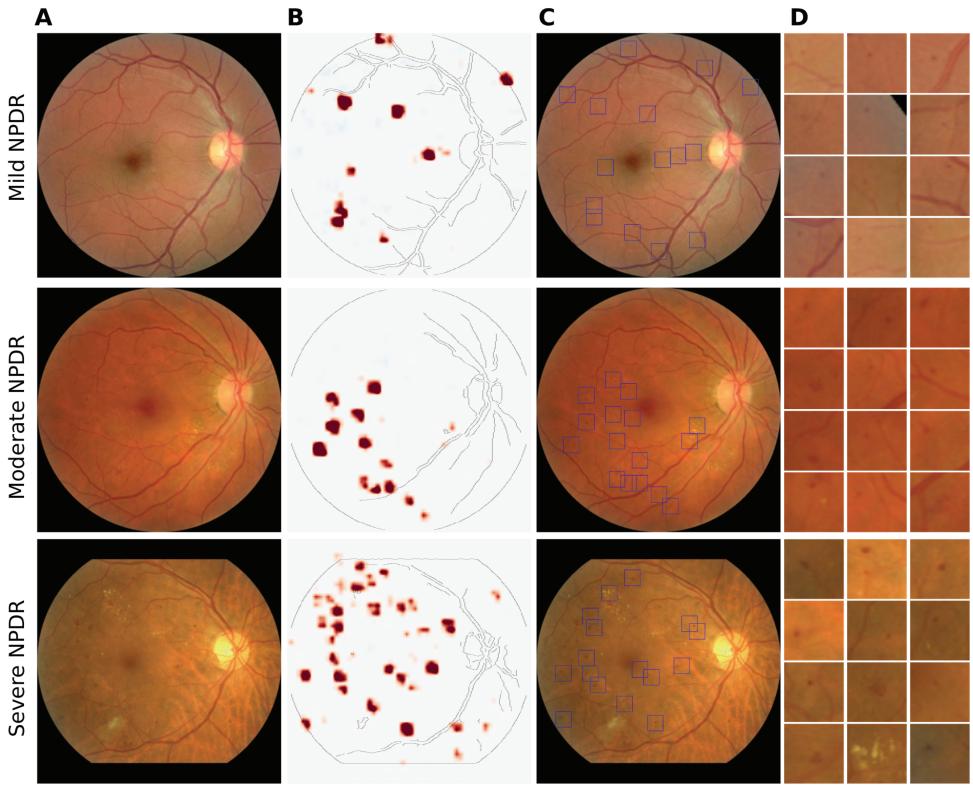

Fig 2. Inherently interpretable deep learning framework highlights clinically relevant image regions. (A) Examples of retinal fundus images from different DR grades (top to bottom: mild NPDR, moderate NPDR and severe NPDR). (B) Class evidence map extracted from the inherently interpretable model without further processing. Red regions indicate evidence for the presence of at least mild DR. (C) Bounding boxes drawn around suspicious regions in the class evidence map. In some cases, the bounding boxes are placed in regions for which there is no visible evidence due to the scaling of the color map. Yet, these evidence values are also strictly positive. (D) Suspicious regions from (C) enlarged and sorted with decreasing evidence scores. Depending on the image grade, the suspicious regions contain various DR-related lesions such as microaneurysms, hemorrhages, or drusen.

https://doi.org/10.1371/journal.pdig.0000831.g002

Although the model was never trained with pixel-level annotations or supervision signals other than the image-level DR reference label, the highlighted regions typically contained DR-related lesions such as microaneurisms, drusen, or hemorrhage with high precision (Fig 3).

We quantitatively evaluated how well the class evidence maps provided information about the location of disease-related lesions using a subset of images from the test set of the development dataset (Fig 3) as well as external datasets with pixel-level annotations (Table 1). The class evidence maps precisely localized DR lesions, as most regions flagged as suspicious indeed contained annotated lesions (Table 2, last column). For the images from the development dataset, we obtained a precision of 0.960 (95% CI [0.941–0.976]), with minor differences between images with mild and moderate NPDR (0.783 vs. 0.970). Notably, our model generalized well to external test sets, with precision ranging from 0.656 to 0.965 (Table 2, last column).

We also evaluated suspicious regions extracted from images the algorithm falsely classified as DR with high confidence (>0.75). To this end, we showed two clinicians 30 images falsely classified as DR with bounding boxes (S8 Fig). Sometimes, these image patches showed

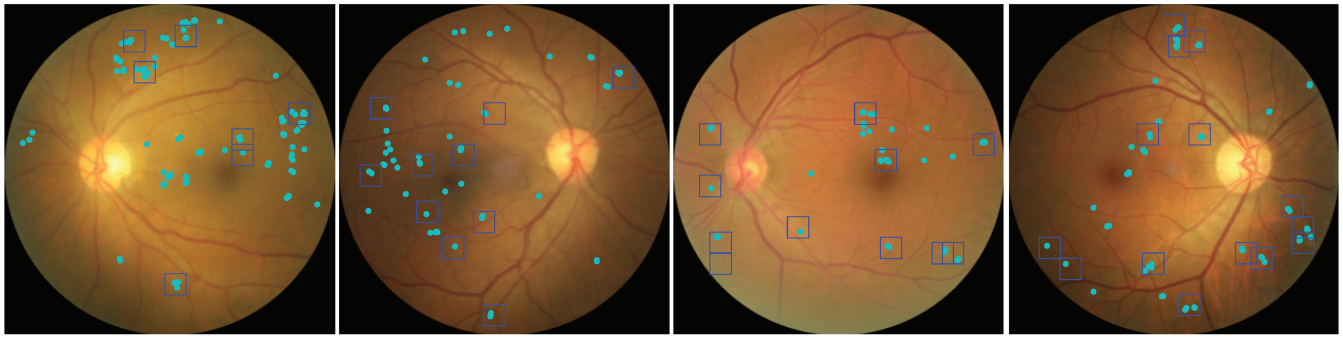

Fig 3. Extracted high evidence images patches contain DR-related lesions. Example fundus images with DR, with DR lesions identified by three clinicians (cyan). Bounding boxes (blue) were extracted from the class evidence maps based on regions of high evidence for DR. Note that all bounding boxes contain annotated lesions, but – as the number of bounding boxes per image was restricted to twelve – not all lesions are contained in bounding boxes.

https://doi.org/10.1371/journal.pdig.0000831.g003

unclear or ambiguous lesions unrelated to DR, but they typically contained anomalies related to DR such as microaneurysms or exudates, but not in a number or severity sufficient for clinical DR diagnosis (S8 Fig).

We next compared the localization performance of the inherently interpretable sparse BagNet to classic post-hoc methods such as Integrated Gradients [38] or Guided Backprop applied to the state-of-the-art model (Fig 4A–4C). These methods were chosen because they performed well in a clinical validation of post-hoc explainability techniques for DR [36]. We found that bounding boxes obtained from Guided Backprop or Integrated Gradients were much less precise in localizing DR-related lesions (0.941 vs. 0.656, Fig 4D, Table 2), especially for out-of-sample test datasets.

We then investigated whether our interpretable deep learning model could effectively aid clinicians in detecting DR via a retrospective reader study with six experienced ophthalmologists screening fundus images for the presence of early DR with various levels of AI assistance (see Methods). Without AI assistance (labeled "H") ophthalmologists reached a mean classification accuracy of 0.611 (95% CI [0.560–0.660]; Fig 5A). Their accuracy increased significantly to 0.758 ([0.711–0.800], $p = 0.0001$, post-hoc test with Tukey's correction for multiple comparisons, see Methods) when they had access to the deep learning model's prediction and confidence ("H+AI"). They achieved similar performance with additional access to AI explanations in the form of bounding boxes around suspicious regions extracted from the class evidence maps ("H+XAI") at an accuracy of 0.786 [0.741–0.825].

We studied ophthalmologists' performance in screening for DR in fundus images of different disease grades in more detail (Fig 5B). Without AI support, detecting images with mild DR (grade 1) was the most challenging with comparably low performance, which improved with AI support. For healthy images, screening performance improved significantly with any form of AI decision support (H: 0.567, [0.477–0.652]; H+AI: 0.842, [0.765–0.897]; H+XAI: 0.817, [0.737–0.876]; H vs. H+AI: $p < 0.0001$; H vs. H+XAI: $p = 0.0001$; H+AI vs. H+XAI: $p = 0.8645$), while for images with mild DR, we observed that screening only improved significantly for AI support with explanations (H: 0.483, [0.395–0.572]; H+AI: 0.617, [0.527–0.699]; H+XAI: 0.733, [0.647–0.805]; H vs. H+AI: $p = 0.0962$; H vs. H+XAI: $p = 0.0003$; H+AI vs. H+XAI: $p = 0.1326$). For images with moderate DR, AI support had no significant effect on screening performance. Taken together, this provides evidence that giving ophthalmologists access to AI support led to superior DR screening performance, with explanations based on the sparse BagNet model being most effective for difficult diagnostic decisions.

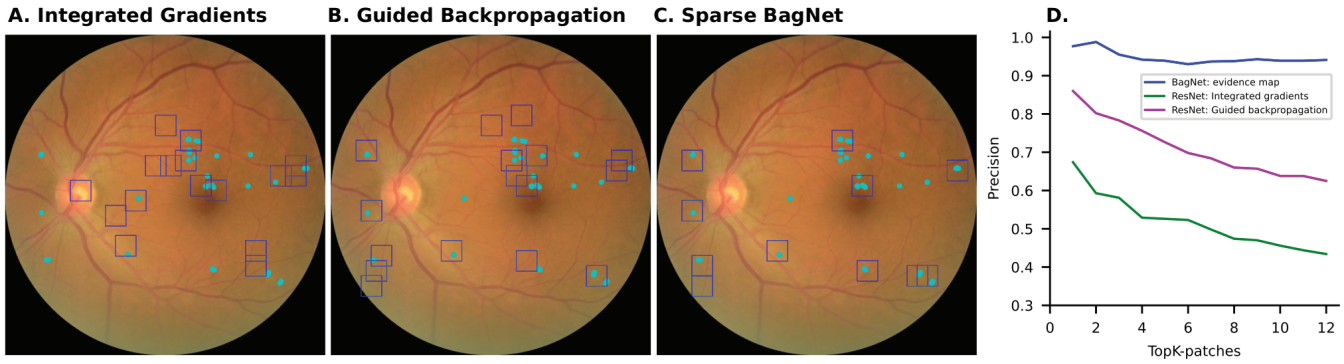

Fig 4. Inherently interpretable deep learning framework highlights lesions more precisely than post-hoc techniques applied to a standard DNN. (A) Suspicious regions (blue) marked with bounding boxes extracted from the heatmap obtained with Integrated gradients from the standard DNN. Clinically relevant DR lesions are marked in cyan. (B) As in (A) extracted from the heatmap obtained with Guided backpropagation. (C) For comparison, suspicious regions were obtained from the SparseBagNet. (D) Systematic comparison of localization precision for clinically annotated DR lesions as a function of the number of considered patches.

https://doi.org/10.1371/journal.pdig.0000831.g004

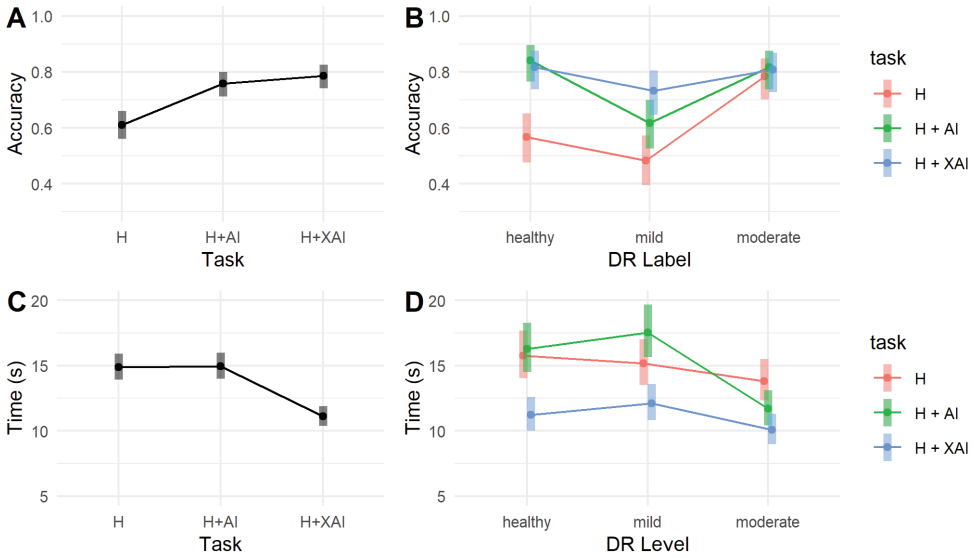

Fig 5. Providing AI-based clinical decision support based on the inherently interpretable deep learning model improves DR screening. (A) Screening accuracy with different levels of AI assistance. Six ophthalmologists graded fundus images without AI assistance ("H"), with access to the AI prediction ("H+AI"), and with additional access to AI explanations ("H+XAI"). AI assistance improved screening accuracy, but access to AI explanations had only a small additional effect. (B) Screening accuracy for DR on fundus images of different disease grades. For healthy images, accuracy improved significantly with any form of AI decision support ("H+AI" or "H+XAI"), while for images with mild DR, screening improved significantly for AI support with explanation ("H+XAI"). For images with moderate DR, AI support had no significant effect on screening performance. (C) Screening time in screening DR with different levels of AI assistance. The decision time is significantly reduced with AI support ("H+XAI") with explanation compared to the other tasks ("H", and "H+AI"). (D) Screening time in screening for DR on fundus images of different disease grades. Screening time reduces at all disease stages with a significant effect of AI decision support with explanation for healthy images ("grade 0"), mild DR ("grade 1"), and moderate DR ("grade 2").

https://doi.org/10.1371/journal.pdig.0000831.g005

We next studied whether AI decision support would not only allow ophthalmologists to make more accurate screening decisions but also reach their decisions faster. We found that

the decision time was significantly reduced when providing ophthalmologists AI support with explanations compared to both other tasks (Fig 5A, H: 15.2 s [14.1-16.4]; H+AI: 15.9 s [14.7-17.1]; H+XAI: 11.7 s [10.8-12.6]; H vs. H+AI: $p = 0.7435$; H vs. H+XAI: $p < 0.0001$; H+AI vs. H+XAI: $p < 0.0001$). This reduction was present at all disease stages, with a significant effect of AI decision support with explanations for healthy images (Fig 5A; H: 15.8 s [14.1-17.7]; H+AI: 16.3 s [14.5-18.3]; H+XAI: 11.2 s [10.0-12.6], H vs. H+AI: $p = 0.9153$; H vs. H+XAI: $p < 0.0001$; H+AI vs. H+XAI: $p < 0.0001$), mild DR (H: 15.2 s [13.5-17.0]; H+AI: 17.5 s [15.6-19.7]; H+XAI: 12.1 s [10.8-13.6], H vs. H+AI: $p = 0.1843$; H vs. H+XAI: $p = 0.180$; H+AI vs. H+XAI: $p < 0.0001$), as well as moderate DR (H: 13.8 s [12.3-15.5]; H+AI: 11.7 s [10.4-13.1]; H+XAI: 10.1 s [9.0-11.3]; H vs. H+AI: $p = 0.1058$; H vs. H+XAI: $p = 0.004$; H+AI vs. H+XAI: $p = 0.1724$). In summary, this indicates that decision support with accurate explanations provided by the sparse BagNet model could reduce screening times across all disease levels.

We also analyzed whether AI decision support would change the confidence with which the ophthalmologists could grade the images, but did not find a significant effect of AI support (H: 3.8 [3.7-3.9]; H+AI: 3.7 [3.6-3.9]; H+XAI: 3.6 [3.5-3.7], H vs. H+AI: $p = 0.6806$; H vs. H+XAI: $p = 0.0543$; H+AI vs. H+XAI: $p = 0.3023$). We conclude that self-reported confidence may not be a reliable measure of grader uncertainty compared to recorded decision time.

We finally analyzed whether the positive effect on accuracy was dependent on whether the deep learning model had classified the image correctly or not, as AI support has been reported to be detrimental in case of model errors [39]. In line with the results above, we found that screening performance and decision time significantly improved for cases in which the deep learning model had made a correct decision (S5 Fig; accuracy, H vs. H+AI: $p<0.0001$; H vs. H+XAI: $p<0.0001$; H+AI vs. H+XAI: $p<0.0001$; time, H vs. H+AI: $p = 0.8178$; H vs. H+XAI: $p<0.0001$; H+AI vs. H+XAI: $p<0.0001$). For cases in which the model had made an incorrect decision, we neither detected positive nor negative effects on accuracy (H vs. H+AI: $p<0.3216$; H vs. H+XAI: $p = 0.4953$; H+AI vs. H+XAI: $p = 0.9480$) and slightly positive effects on decision time (H vs. H+AI: $p = 0.4557$; H vs. H+XAI: $p = 0.0941$; H+AI vs. H+XAI: $p = 0.0031$) meaning that the decision time was still smaller despite the wrong prediction of the model.

Discussion

In this study, we trained and evaluated an inherently interpretable deep-learning model for early diabetic retinopathy detection. This is a challenging task even for experienced ophthalmologists. Our model achieved a classification performance comparable to the black-box baseline model in the internal test set and on ten publicly available external datasets. While the training dataset contained a large fraction of images from patients of Latin American ethnicity, the external datasets were acquired in diverse world regions and different devices, thus that our model showed a good generalization across different ethnic groups and patient populations. While some of these datasets also contained patients of African ancestry, none of the datasets were acquired on the African continent.

In addition to a binary diagnostic decision that is commonly communicated in DR screening settings, our model provides explanations via interpretable evidence maps, which highlight regions of the image used by the network in making its decisions. We found that the inherently interpretable framework precisely located disease-related lesions in the image,

more so than post-hoc techniques applied to a state-of-the-art DNN, in particular for out-of-sample test datasets. Even in case of incorrect model predictions according to the reference labels, model explanations proofed to be useful and highlighted suspicious regions.

In a retrospective reader study, we found that highlighting these regions during grading helped ophthalmologists improve their grading performance, especially for difficult cases, while reducing their decision time. This indicates that current paradigms used in AI-based screening scenarios may benefit from including explanations for easier human verification and enhanced trust in the algorithms decision [3,5]. Our study further showed that the errors of the AI model did not negatively affect decision-making by ophthalmologists, in contrast to earlier human-AI studies on clinical decision support [39,40]. A limitation of our model is that it was trained on a dataset from North America, and may need to be fine-tuned on data from the intended target population, although its generalization results on ten additional datasets were promising.

As the potential of AI for medical image analysis has become evident [41,42], such systems have reached performance close to, or even superior to, those of clinical experts in a variety of tasks [43]. More recently, the focus has shifted towards AI systems assisting clinicians in making better decisions [39]. In this setting, clinicians need to understand how decisions are formed by the AI model, such that transparency and interpretability of medical AI systems have become important aspects [7,11,12,44]. In agreement, the need for trustworthy and transparent AI systems and effective human/AI collaboration has been identified in standardized guidelines to facilitate their adoption in clinical practice [44]. While this generally poses challenges in balancing high performance and interpretability [45], our study has shown that inherent interpretability can be achieved without significant performance trade-offs if the inductive biases of the interpretable model are met – in our case, as early DR causes only very localized lesions in the retina. Other inherently interpretable models include prototype-based networks [46], which are difficult to use for diseases with many small, distributed lesions, for which the training procedure is more complex and for which interpretability is not straightforward [47].

In a clinical setting, such an inherently interpretable model could assist clinicians in mitigating the challenge of early and accurate diagnosis of presymptomatic diseases, such as diabetic retinopathy detection. Given several approved AI systems for DR screening, clinical implementation could be comparatively straightforward. The trained model can efficiently generate predictions, on a time scale not impeding on clinical practice ($\ll 1s$/image), requires relatively little memory ($\sim 350mb$), and does not require additional models to run to create explanations. Such explanations could be added to existing reporting templates in commercial AI systems, allowing screeners to quickly ascertain the plausibility of the models prediction. In this setting, real-world prospective studies could be conducted to test the impact of the explanations obtained from our model on screening quality and speed, in particular for patient with beginning DR.

One limitation of our model is that it may not provide good explanations if its inductive bias is not matched to the disease, e.g. when lesions cover large parts of the retina as in more advanced DR grades [13]. Future applications also include time-to-progression prediction for diseases like DR [48] through interpretable-by-design deep survival models [49].

Supporting information

S1 Fig. Web interface for the annotation task. A fundus image is shown and based on it, the annotator is asked to annotate lesions related to Diabetic Retinopathy. By moving the mouse over a region of the image, an enlarged version of that region is displayed. All images are from

patients with DR of grade 1 ("mild DR") or 2 ("moderate DR"). Each lesion is marked by selecting the type (Microaneurysms: MA, hemorrhages: HE, exudates: EX, soft exudate: SE, artifact, or any other lesions) and clicking on the image location.
(TIF)

S1 Table. **Summary of model performance on localizing DR-related lesions from graders' annotations** The precision of the model on each clinician annotation is calculated as the proportion of bounding boxes from regions highlighted on heatmaps containing lesions annotated by a grader. The random precision is obtained by drawing 20 random bounding boxes over each annotated image, excluding those falling in regions containing more than 10% black pixels. The union "∪" gives the precision of the model with the combined clinicians' annotation masks, while the intersection "∩" gives the precision of the model with reference annotation masks obtained as the intersections of clinicians' annotation over each image.
(PDF)

S2 Table. **Inter-grader performance on 65 fundus images from the internal Kaggle test set annotated by three ophthalmologists** "Grader X - Grader Y" refers to the dice score between grader X and grader Y. The Dice score is calculated for each pair of graders as the overlap between their annotation using a patch size of 33×33 pixels corresponding to the receptive field of the model and considering different strides (s = 8, 32 for overlapping patches and s=33 for non-overlapping patches). "Grader X - Grader Y ∪ Grader Z" refers to the dice score between grader X, Y, and Z while "Grader Y ∪ Grader Z" is the union between grader Y and Z, and "Grader Y ∩ Grader Z" is the intersection between grader Y and Z.
(PDF)

S2 Fig. **Comparison of the sparse BagNet performance with different regularization values on the validation dataset** The regularization coefficient λ affects the classification performance (accuracy and AUC) of the model. The red points indicate the selected value, which is a compromise between sparsity and both accuracy and AUC. It also defines the trade-off between the model's interpretability and classification performance.
(TIF)

S3 Fig. **Web interface for the grading task without AI support ("H")** A fundus image is shown and based on it, the grader is asked to decide whether the corresponding patient has Diabetic Retinopathy (DR) of any severity, including mild DR. In addition, the grader is asked to rate the confidence of his/her decision on a scale from 1 (least confident) to 5 (most confident). By moving the mouse over a region of the image, an enlarged version of that region is displayed. The time taken to reach each decision (grading and confidence) is recorded.
(TIF)

S4 Fig. **Web interface for the grading task with AI support ("H + AI")** A fundus image is shown with the model's prediction and its confidence level (from 0% to 100 %, with 100% being the highest confidence score). Based on this, the grader is asked to decide whether the corresponding patient has Diabetic Retinopathy (DR) of any severity, including mild DR. In addition, the grader is asked to rate the confidence of his/her decision on a scale from 1 (least confident) to 5 (most confident). By moving the mouse over a region of the image, an enlarged version of that region is displayed. The time taken to reach each decision (grading and confidence) is recorded.
(TIF)

S5 Fig. **Web interface for the grading task with AI support and explanations ("H + XAI").** A fundus image is shown with the model's prediction, its confidence level (from 0% to 100 %,

with 100% being the highest confidence score), and explanation in the form of blue bounding boxes around the regions for which the AI model believes that they contain signs of DR. Based on this, the grader is asked to decide whether the corresponding patient has Diabetic Retinopathy (DR) of any severity, including mild DR. In addition, the grader is asked to rate the confidence of his/her decision on a scale from 1 (least confident) to 5 (most confident). By moving the mouse over a region of the image, an enlarged version of that region is displayed. The time taken to reach each decision (grading and confidence) is recorded.
(TIF)

S6 Fig. Examples of fundus images from each dataset.
(TIF)

S7 Fig. Heatmap with combined clinicians' annotations of four examples of fundus cases with DR. For each example, the left side shows the heatmap with clinicians' annotations and bounding boxes around the regions of positive activation, while the right side shows the fundus image with clinicians' annotations and bounding boxes around the regions of positive activation.
(TIF)

S8 Fig. Examples of high-confidence false positives analyzed by two clinicians. On the left side of each example, the image displays bounding boxes highlighting regions with positive activation. On the right side, the suspicious regions from the left are enlarged and arranged in descending order of evidence scores. (A) A false-positive image where the clinicians interpreted the suspicious regions as "vitreous opacities" and "uveitis vitreous cells," respectively. (B) A false-positive image where one clinician identified the suspicious regions as "synchisis scintillans," while the other suggested the patient may have recently received an intravitreal steroid injection. (C) A false-positive image where both clinicians identified the suspicious regions as "microaneurysms" possibly associated with bleeding. (D) A false-positive image where both clinicians recognized the suspicious regions as "microaneurysms" and "hard exudates". **(E, F)** False-positive images where one clinician classified the image as DR while the other classifies it as no DR, citing the presence of only a single microaneurysm lesion in the suspicious regions.
(TIF)

S9 Fig. Analysis of errors of the AI model on accuracy and decision times for different tasks during the retrospective reader study. (a) For all tasks, ophthalmologists' accuracy is higher when the deep learning model makes the correct decision. For correct classifications, the AI assistance improves grading accuracy. For incorrect classification, it does not make it worse. **(b)** Ophthalmologists' decision time decreases overall when the deep learning model makes the correct decision. When the AI model is correct, the explanation decreases decision time significantly, while it does not increase the decision time for incorrect decisions.
(TIF)

Acknowledgments

We thank Sarah Müller, Pearse Keane and Murat Ayhan for discussion and Murat Ayhan quality filtering code.

Author contributions

Conceptualization: Kerol Djoumessi, Lisa M. Koch, Philipp Berens.

Data curation: Kerol Djoumessi, Ziwei Huang, Annekatrin Rickmann, Natalia Simon, Lisa M. Koch.

Formal analysis: Kerol Djoumessi.

Funding acquisition: Philipp Berens, Lisa M. Koch.

Investigation: Kerol Djoumessi, Laura Kühlewein, Lisa M. Koch, Philipp Berens.

Methodology: Kerol Djoumessi, Philipp Berens.

Project administration: Philipp Berens.

Software: Kerol Djoumessi, Ziwei Huang.

Supervision: Lisa M. Koch, Philipp Berens.

Validation: Laura Kühlewein, Annekatrin Rickmann, Natalia Simon.

Visualization: Kerol Djoumessi.

Writing – original draft: Kerol Djoumessi, Lisa M. Koch, Philipp Berens.

Writing – review & editing: Laura Kühlewein, Annekatrin Rickmann, Natalia Simon.

References

1. Gulshan V, Peng L, Coram M, Stumpe MC, Wu D, Narayanaswamy A, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA. 2016;316(22):2402–10. https://doi.org/10.1001/jama.2016.17216 PMID: 27898976
2. Food US, Administration D. Artificial Intelligence and Machine Learning (AI/ML)-Enabled Medical Devices (SaMD) Action Plan; 2021. Available from: https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-aiml-enabled-medical-devices
3. Grzybowski A, Brona P, Lim G, Ruamviboonsuk P, Tan GSW, Abramoff M, et al. Artificial intelligence for diabetic retinopathy screening: a review. Eye (Lond). 2020;34(3):451–60. https://doi.org/10.1038/s41433-019-0566-0 PMID: 31488886
4. Ipp E, Liljenquist D, Bode B, Shah VN, Silverstein S, Regillo CD, et al. Pivotal evaluation of an artificial intelligence system for autonomous detection of referrable and vision-threatening diabetic retinopathy. JAMA Netw Open. 2021;4(11):e2134254. https://doi.org/10.1001/jamanetworkopen.2021.34254 PMID: 34779843
5. Grauslund J. Diabetic retinopathy screening in the emerging era of artificial intelligence. Diabetologia. 2022;65(9):1415–23. https://doi.org/10.1007/s00125-022-05727-0 PMID: 35639120
6. Grote T, Berens P. On the ethics of algorithmic decision-making in healthcare. J Med Ethics. 2020;46(3):205–11. https://doi.org/10.1136/medethics-2019-105586 PMID: 31748206
7. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell. 2019;1(5):206–15. https://doi.org/10.1038/s42256-019-0048-x PMID: 35603010
8. Chetoui M, Akhloufi MA. Explainable diabetic retinopathy using EfficientNET. Annu Int Conf IEEE Eng Med Biol Soc. 2020;2020:1966–9. https://doi.org/10.1109/EMBC44109.2020.9175664 PMID: 33018388
9. Alghamdi HS. Towards explainable deep neural networks for the automatic detection of diabetic retinopathy. Appl Sci. 2022;12(19):9435. https://doi.org/10.3390/app12199435
10. Gonzalez-Gonzalo C, Liefers B, van Ginneken B, Sanchez CI. Iterative augmentation of visual evidence for weakly-supervised lesion localization in deep interpretability frameworks: application to color fundus images. IEEE Trans Med Imaging. 2020;39(11):3499–511. https://doi.org/10.1109/TMI.2020.2994463 PMID: 32746093
11. Ghassemi M, Oakden-Rayner L, Beam AL. The false hope of current approaches to explainable artificial intelligence in health care. Lancet Digit Health. 2021;3(11):e745–50. https://doi.org/10.1016/S2589-7500(21)00208-9 PMID: 34711379

12. Grote T, Berens P. How competitors become collaborators-Bridging the gap(s) between machine learning algorithms and clinicians. Bioethics. 2022;36(2):134–42. https://doi.org/10.1111/bioe.12957 PMID: 34599834
13. Kerol D, Ilanchezian I, Kühlewein L, Faber H, Baumgartner C, Bah B, et al. Sparse activations for interpretable disease grading. Med Imaging Deep Learn. 2023.
14. Brendel W, Bethge M. Approximating CNNs with Bag-of-local-Features models works surprisingly well on ImageNet. 2019:1–10. https://doi.org/10.1109/ICLR.2019.00001
15. Solomon SD, Chew E, Duh EJ, Sobrin L, Sun JK, VanderBeek BL, et al. Diabetic retinopathy: a position statement by the american diabetes association. Diabetes Care. 2017;40(3):412–8. https://doi.org/10.2337/dc16-2641 PMID: 28223445
16. Vujosevic S, Aldington SJ, Silva P, Hernández C, Scanlon P, Peto T, et al. Screening for diabetic retinopathy: new perspectives and challenges. Lancet Diabetes Endocrinol. 2020;8(4):337–47. https://doi.org/10.1016/S2213-8587(19)30411-5 PMID: 32113513
17. Wilkinson CP, Ferris FL 3rd, Klein RE, Lee PP, Agardh CD, Davis M, et al. Proposed international clinical diabetic retinopathy and diabetic macular edema disease severity scales. Ophthalmology. 2003;110(9):1677–82. https://doi.org/10.1016/S0161-6420(03)00475-5 PMID: 13129861
18. Dugas E, Jared J, Cukierski W. Diabetic Retinopathy Detection; 2015. Available from: https://kaggle.com/competitions/diabetic-retinopathy-detection
19. Porwal P, Pachade S, Kamble R, Kokare M, Deshmukh G, Sahasrabuddhe V, et al. Indian Diabetic Retinopathy Image Dataset (IDRiD): a database for diabetic retinopathy screening research. Data. 2018;3(3):25. https://doi.org/10.3390/data3030025
20. Decencière E, Cazuguel G, Zhang X, Thibault G, Klein J-C, Meyer F, et al. TeleOphta: machine learning and image processing methods for teleophthalmology. IRBM. 2013;34(2):196–203. https://doi.org/10.1016/j.irbm.2013.01.010
21. Zhou Y, Wang B, Huang L, Cui S, Shao L. A benchmark for studying diabetic retinopathy: segmentation, grading, and transferability. IEEE Trans Med Imaging. 2021;40(3):818–28. https://doi.org/10.1109/TMI.2020.3037771 PMID: 33180722
22. Kauppi T, Kalesnykiene V, Kamarainen JK, Lensu L, Sorri I, Raninen A, et al. The diaretdb1 diabetic retinopathy database and evaluation protocol. BMVC. 2007;1:10.
23. Li T, Gao Y, Wang K, Guo S, Liu H, Kang H. Diagnostic assessment of deep learning algorithms for diabetic retinopathy screening. Inf Sci. 2019;501:511–22.
24. Pires R, Jelinek HF, Wainer J, Valle E, Rocha A. Advancing bag-of-visual-words representations for lesion classification in retinal images. PLoS One. 2014;9(6):e96814. https://doi.org/10.1371/journal.pone.0096814 PMID: 24886780
25. Karthik SD Maggie. APTOS 2019 blindness detection; 2019. Available from: https://kaggle.com/competitions/aptos2019-blindness-detection
26. Benítez-VEC MI, Román J, Noguera J, García-Torres M, Ayala J. Dataset from fundus images for the study of diabetic retinopathy. Data Brief. 2021;36:107068. https://doi.org/10.1016/j.dib.2021.107068
27. Decencière E, Zhang X, Cazuguel G, Lay B, Cochener B, Trone C, et al. Feedback on a publicly distributed image database: the messidor database. Image Anal Stereol. 2014;33(3):231. https://doi.org/10.5566/ias.1155
28. Abràmoff MD, Folk JC, Han DP, Walker JD, Williams DF, Russell SR, et al. Automated analysis of retinal images for detection of referable diabetic retinopathy. JAMA Ophthalmol. 2013;131(3):351–7. https://doi.org/10.1001/jamaophthalmol.2013.1743 PMID: 23494039
29. Koch LM, Baumgartner CF, Berens P. Distribution shift detection for the postmarket surveillance of medical AI algorithms: a retrospective simulation study. NPJ Digit Med. 2024;7(1):120. https://doi.org/10.1038/s41746-024-01085-w PMID: 38724581
30. Tan M, Le Q. Efficientnet: Rethinking model scaling for convolutional neural networks. International conference on machine learning. 2019;90:6105–14.
31. Liu R, Wang X, Wu Q, Dai L, Fang X, Yan T, et al. DeepDRiD: diabetic retinopathy-grading and image quality estimation challenge. Patterns (N Y). 2022;3(6):100512. https://doi.org/10.1016/j.patter.2022.100512 PMID: 35755875
32. Boreiko V, Ilanchezian I, Ayhan M, Muller S, Koch L, Faber H. Visual explanations for the detection of diabetic retinopathy from retinal fundus images. In: Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer; 2022. p. 539–49.
33. Mueller S, Heidrich H, Koch LM, Berens P. Fundus circle cropping. Available from: https://github.com/berenslab/fundus_circle_cropping
34. Huang Y, Lin L, Cheng P, Lyu J, Tang X. Identifying the key components in ResNet-50 for diabetic retinopathy grading from fundus images: a systematic investigation. arXiv preprint 2021. https://arxiv.org/abs/2110.14160

35. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR); 2016.
36. Ayhan MS, Kümmerle LB, Kühlewein L, Inhoffen W, Aliyeva G, Ziemssen F, et al. Clinical validation of saliency maps for understanding deep neural networks in ophthalmology. Med Image Anal. 2022;77:102364. https://doi.org/10.1016/j.media.2022.102364 PMID: 35101727
37. Ferrer L, Riera P. Confidence Intervals for evaluation in machine learning. Available from: https://github.com/luferrer/ConfidenceIntervals
38. Sundararajan M, Taly A, Yan Q. Axiomatic attribution for deep networks. In: Proceedings of the International Conference on Machine Learning. PMLR; 2017. p. 3319–28.
39. Tschandl P, Rinner C, Apalla Z, Argenziano G, Codella N, Halpern A, et al. Human-computer collaboration for skin cancer recognition. Nat Med. 2020;26(8):1229–34. https://doi.org/10.1038/s41591-020-0942-0 PMID: 32572267
40. Ng AY, Oberije CJG, Ambrózay É, Szabó E, Serfőző O, Karpati E, et al. Prospective implementation of AI-assisted screen reading to improve early detection of breast cancer. Nat Med. 2023;29(12):3044–9. https://doi.org/10.1038/s41591-023-02625-9 PMID: 37973948
41. Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, et al. A survey on deep learning in medical image analysis. Med Image Anal. 2017;42:60–88. https://doi.org/10.1016/j.media.2017.07.005 PMID: 28778026
42. Muehlematter UJ, Daniore P, Vokinger KN. Approval of artificial intelligence and machine learning-based medical devices in the USA and Europe (2015–2020): a comparative analysis. Lancet Digit Health. 2021;3(3):e195–203. https://doi.org/10.1016/S2589-7500(20)30292-2 PMID: 33478929
43. Liu X, Faes L, Kale AU, Wagner SK, Fu DJ, Bruynseels A, et al. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. Lancet Digit Health. 2019;1(6):e271–97. https://doi.org/10.1016/S2589-7500(19)30123-2 PMID: 33323251
44. González-Gonzalo C, Thee EF, Klaver CCW, Lee AY, Schlingemann RO, Tufail A, et al. Trustworthy AI: Closing the gap between development and integration of AI systems in ophthalmic practice. Prog Retin Eye Res. 2022;90:101034. https://doi.org/10.1016/j.preteyeres.2021.101034 PMID: 34902546
45. Frasca M, La Torre D, Pravettoni G, Cutica I. Explainable and interpretable artificial intelligence in medicine: a systematic bibliometric review. Discov Artif Intell. 2024;4(1):15. https://doi.org/10.1007/s44163-024-00114-7
46. Chen C, Li O, Tao D, Barnett A, Rudin C, Su JK. This looks like that: deep learning for interpretable image recognition. Adv Neural Inf Process Syst. 2019:32.
47. Djoumessi K, Bah B, Kühlewein L, Berens P, Koch L. This actually looks like that: Proto-BagNets for local and global interpretability-by-design. Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer; 2024. p. 718–28.
48. Dai L, Sheng B, Chen T, Wu Q, Liu R, Cai C, et al. A deep learning system for predicting time to progression of diabetic retinopathy. Nat Med. 2024;30(2):584–94. https://doi.org/10.1038/s41591-023-02702-z PMID: 38177850
49. Gervelmeyer J, Mueller S, Djoumessi K, Merle D, Clark S, Koch L. Interpretable-by-design deep survival analysis for disease progression modeling. In:International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer; 2024. p. 502–12.

This actually looks like that: Proto-BagNets for local and global interpretability-by-design

Kerol Djoumessi[1,2(✉)][0009−0004−1548−9758], Bubacarr Bah[3], Laura Kühlewein[4], Philipp Berens[1,2(✉)][0000−0002−0199−4727], and Lisa Koch[1,2,5(✉)][0000−0003−4377−7074]

[1] Hertie Institute for AI in Brain Health, University of Tübingen, Germany
{kerol.djoumessi-donteu, philipp.berens}@uni-tuebingen.de
[2] Tübingen AI Center, University of Tübingen, Germany
[3] Medical Research Council Unit The Gambia at London School of Hygiene and Tropical Medicine
[4] University Eye Clinic, University of Tübingen, Germany
[5] Department of Diabetes, Endocrinology, Nutritional Medicine and Metabolism UDEM, Inselspital, Bern University Hospital, University of Bern, Switzerland
{lisa.koch}@unibe.ch

Abstract. Interpretability is a key requirement for the use of machine learning models in high-stakes applications, including medical diagnosis. Explaining black-box models mostly relies on post-hoc methods that do not faithfully reflect the model's behavior. As a remedy, prototype-based networks have been proposed, but their interpretability is limited as they have been shown to provide coarse, unreliable, and imprecise explanations. In this work, we introduce Proto-BagNets[6], an interpretable-by-design prototype-based model that combines the advantages of bag-of-local feature models and prototype learning to provide meaningful, coherent, and relevant prototypical parts needed for accurate and interpretable image classification tasks. We evaluated the Proto-BagNet for drusen detection on publicly available retinal OCT data. The Proto-BagNet performed comparably to the state-of-the-art interpretable and non-interpretable models while providing faithful, accurate, and clinically meaningful local and global explanations.

Keywords: Interpretability-by-design · Optical Coherence Tomography · Part-prototype networks

1 Introduction

For adopting deep learning models in safety-critical applications such as medical diagnosis, it is crucial that users can understand why a model produced a specific output [16]. This form of interpretability is usually obtained either through post-hoc explanations of black-box models [20] or through architectural design [4, 13]. Post-hoc methods [17, 20] interpret an approximation of the

[6] Code available at https://github.com/kdjoumessi/Proto-BagNets

true decision mechanism [11] through saliency maps. These highlight the most discriminating regions in the input, but often provide inaccurate and unfaithful explanations [1, 16]. To remedy this, several approaches have been proposed with structurally built-in interpretability, such as bag-of-local-features models (BagNets) [3], concept-based models [15], and prototype-based models [4].

The BagNet [3] is an implicitly patch-based interpretable-by-design model with a small receptive field, where predictions solely rely on local evidence. Its recent modification [13] provides sparse and fine-grained local class activation maps, but does not allow humans to gain a global understanding of the model's decision. Concept-based models [15] follow a case-based reasoning process where high-level representations of the data (concepts) are learned and used to classify new images. Prototype-based networks [4] can be seen as a special case of concept-based models, in which learned concepts are replaced by the representative training image parts (prototypes) to improve interpretability. In these models, similarities to the learned prototypes are used to classify new examples. Explanations can be obtained during inference by highlighting, for a query image, its prototypical parts most similar to each learned prototype, thus providing both local explanations thanks to the similarity map and global explanations through the visualization of the learned prototypes. ProtoPNet [4], the first prototype-based network, has gained considerable attention due to its easy-to-understand architecture and high-level reasoning process close to that of humans in solving complex tasks. Although numerous variants have been proposed to improve its performance and interpretability [2, 7, 10, 14], applications in medical imaging remain relatively limited. This may in part be due to the fact that the interpretability of prototype-based models is more limited than appears at first glance, as it has been shown that they do not actually provide faithful explanations [7, 19].

We propose Proto-BagNet, an interpretable-by-design prototype-based model that combines the local and fine-grained interpretability of BagNet with the global interpretability of prototype learning. We integrated recent advances in training prototype-based models and proposed an additional prototype diversity constraint. We evaluated our model for detecting drusen lesions on Optical Coherence Tomography (OCT) images and showed that the Proto-BagNet preserves high predictive performance while providing faithful, clinically meaningful, and precise explanations. Our model explanations accurately localized drusen both in the learned prototypes and query test images.

2 Developing a faithful prototype-based network

2.1 Baseline ProtoPNet model

We built on the ProtoPNet [4] as a baseline, which consists of three main components: a backbone feature extractor f, a prototype layer g_p, and a classification layer h. Given an input image $\mathbf{X} \in \mathbb{R}^{H \times W \times C}$ (with height H, width W, and number of channels C), the backbone first extracts a meaningful feature representation $\mathbf{Z} = f(\mathbf{X}) \in \mathbb{R}^{M \times N \times D}$. The prototype layer $g_{\mathbf{p}}$ consists of $b = m \times c$

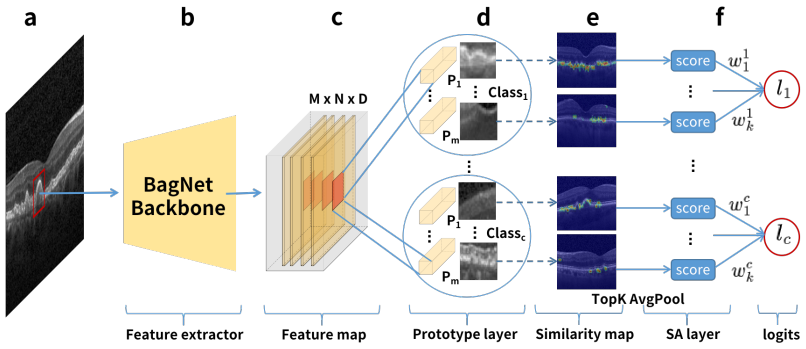

Fig. 1. Architecture of the Proto-BagNet. **(a)** Example OCT B-scan image. The red patch illustrates the small receptive field of **(b)** the BagNet backbone. **(c)** Feature map and **(d)** Prototype layer with m prototypes per class. **(e)** Resulting similarity maps from each prototype to the input. **(f)** The soft aggregation layer aggregates the average top-k scores from each similarity map into their allocated categories for classification.

learnable prototypes $\mathbf{P} = \{\mathbf{p}_j^i \in \mathbb{R}^{H_P \times W_P \times D}\}_{j=1}^m$ (typically $H_P = W_P = 1$) where m denotes the number of prototypes per class, c the number of classes, and $\mathbf{p}_j^i$ the j-th prototype of class i. For each prototype $\mathbf{p}_j^i$, the prototype layer computes a similarity map $M_{\mathbf{p}_j^i}^{\mathbf{x}} = \text{Sim}(\mathbf{Z}, \mathbf{p}_j^i) \in \mathbb{R}^{M \times N}$. The similarity score between a prototype $\mathbf{p}_j^i$ and the feature vector $\mathbf{z}^{(h,w)} \in \mathbf{Z}$ is defined as $s_{i,j}^{(h,w)} = \log\left((d_{i,j}^{h,w} + 1)/(d_{i,j}^{h,w} + \epsilon)\right)$, where $d_{i,j}^{h,w} = \|\mathbf{p}_j^i - \mathbf{z}^{(h,w)}\|_2^2$. The similarity maps contain positive scores indicating where and to what extent prototypes are present in an image. ProtoPNet uses the highest value of the similarity map $g_{\mathbf{p}_j^i} = \max(M_{\mathbf{p}_j^i}^{\mathbf{x}})$ as the final similarity score between $\mathbf{p}_j^i$ and $\mathbf{X}$, indicating how strong the prototype $\mathbf{p}_j^i$ is present in $\mathbf{X}$. Finally, the b similarity scores from the prototype layer g_p are aggregated in the fully connected layer h to generate the final classification logits. To make the prototypes visualizable as specific prototypical parts of a sample, the learned prototypes are replaced with the closest feature representation from real training images to ensure interpretability.

ProtoPNet explains its predictions for a given image ("local explanation") by (1) visualizing the similarity map for each prototype $\mathbf{p}_j^i$ and (2) by computing the smallest bounding box enclosing the $95th$ percentile of all similarity values [4], providing the corresponding cropped region as the most similar part to the learned prototype to imply *'this part of the input looks like that learned prototype'*. The same approach is used to provide explanations of the concepts learned by the model ("global explanations") by cropping the prototypes from the most similar training image. However, ProtoPNet provides only coarse-grained similarity maps due to the large receptive field size of the model [7]. Furthermore, the explanation is not faithful to the model, as the cropped area does not correspond to the model's receptive field. As a result, ProtoPNet provides both imprecise

local and global explanations of its decisions. These issues are likely to be shared by all prototype-based models derived from ProtoPNet [19].

2.2 Enhancing interpretability with the BagNet backbone

As prototypes are learned in the feature space, the backbone feature extractor plays a crucial role for interpretability. It implicitly determines the size of the learned prototypes through its receptive field and thus the size of the explanation. The ProtoPNet and its variants use classical architectures such as ResNet-50 [4, 10, 14], resulting in large receptive fields (e.g. 427×427 for a ResNet-50 backbone) with variable explanation size. Here, we propose to replace the feature extractor with a BagNet architecture [3, 13] (Fig. 1b), leading to a model we called Proto-BagNet (Fig. 1). The feature map $\mathbf{Z} = f(\mathbf{X})$ is extracted from the BagNet's penultimate layer (Fig. 1c, typically $D = 2048$). This architecture leads to a small fixed receptive field and prototypes of size $r \times r$ independent of the input size. It also allows for higher-resolution feature and prototype similarity maps, and can therefore provide localized, fine-grained explanations.

2.3 Integrating recent advances in training prototype-based models

In addition, we implemented recent advances in prototype-based networks training [2, 10, 14]. (1) To prevent prototypes of one class from contributing to the prediction of other classes, we replaced the fully connected classification layer of ProtoPNet with a soft aggregation (SA) layer (Fig. 1f) [10], which aggregates the prototypes' similarity scores only in their assigned classes, setting weights between classes to zero. (2) To enable the model to consider multiple image regions for the classification task instead of considering only the region with the highest score of each similarity map as in ProtoPNet, we considered the top-k scores through average pooling as $g_{\mathbf{p}j^i} = \text{AvgPool}\big(\text{topk}(M^{i}_{\mathbf{p}_j})\big)$ [2]. Thus, our similarity map indicates to what extent a prototype is present on average in the k most similar prototypical parts of the input. (3) We regularized the prototype layer by adding a sparsity constraint to each similarity map as in [13,14], to constrain activation to discriminative input regions. (4) Finally, as we noticed redundant prototypes (often extracted from the same training image, see Suppl. Fig. 1), we introduced a dissimilarity loss (see below) to prevent the network from learning duplicate prototypes while promoting their coherence and uniqueness. Thus, the total loss function was:

$$\mathcal{L} = \mathcal{L}_{ce} + \underbrace{\lambda_{clst}\mathcal{L}_{clst} + \lambda_{sep}\mathcal{L}_{sep}}_{\text{ProtoPNet}} + \lambda_{L1,c}\mathcal{L}_{L1,c} + \lambda_{L1,s}\mathcal{L}_{L1,s} - \lambda_{diss} \sum_{p_i, p_j} \|\mathbf{p}_i - \mathbf{p}_j\|^2$$

Here, $\mathcal{L}_{\text{ce}}$ is the cross-entropy loss; $\mathcal{L}_{clst}$ and $\mathcal{L}_{sep}$ the cluster and separation losses from ProtoPNet [4], $\mathcal{L}_{L1,c}$ is the ℓ_1 regularization of the classification layer as in [4]; $\mathcal{L}_{L1,s}$ regularizes the similarity maps [13,14]. Finally, $\sum_{p_i,p_j} \|\mathbf{p}_i - \mathbf{p}_j\|^2$ is our proposed dissimilarity loss with $\mathbf{p}_i, \mathbf{p}_j \in \mathbf{P}, i \neq j$.

Our architectural changes alongside with the modified loss function led to an interpretable-by-design prototype-based model (Proto-BagNet) that is easier to interpret, relying on the small receptive field of the model for predictions and explanations. In addition, Proto-BagNet provides accurate local and global explanations (see Sec. 3.3, and 3.4) in the form of *'this part of the input actually looks like that learned prototype'*.

3 Results

3.1 Dataset

We used a publicly available, anonymized dataset [12] consisting of retinal OCT B-scans from patients with various diseases (drusen, DME, CNV). We focused on the binary task of drusen detection and filtered out images with DME and CNV diagnoses, as well as low-resolution images (width < 496). To counter the class imbalance, we removed half of the healthy images, leading to a dataset of $34,962$ images ($8,616$ drusen, $26,346$ healthy). We split the resulting dataset into training (80%) and validation (20%) sets, preserving the imbalance proportion (73% vs 27%) and ensuring that all B-scans from each patient were assigned to the same set. We then used the separate test set included in the dataset for evaluation, consisting of 250 healthy and 248 drusen images (51% vs 49%), reflecting the high variability of drusen prevalence according to age group [18]. All images were resized to 496×496 and normalized by the mean and standard deviation of the training set. To evaluate the relevance of the learned prototypes, an experienced in-house ophthalmologist provided detailed drusen annotations on a selection of 40 test images.

3.2 Proto-BagNet yields good accuracy on drusen detection

We first evaluated the classification performance of our method for a clinically relevant binary task of detecting patient's OCT-B scans with drusen lesions (lipid deposits under the retina [6]), characteristic of age-related macular degeneration and diabetic retinopathy [6, 12]. For Proto-BagNet, we configured the backbone feature extractor (BagNet model) to a receptive field size $r = 33$ as in [13]. Hyperparameters including regularization coefficients, data augmentation, and the number of prototypes were optimized on the validation dataset using a grid search, while λ_{sep} and λ_{clst} were set as in [4]. Based on the validation performance, we set $k = 5$ considering the average top-5 and used $m = 5$ prototypes per class, which lead to a total of $b = 10$ prototypes.

We compared Proto-BagNet against ProtoPNet with a ResNet-50 backbone [4] and with non-prototype classification networks such as a dense BagNet [13] and ResNet-50 [9]. We followed the same training procedure for ProtoPNet (with $\lambda_{L1,s} = \lambda_{diss} = 0$, $K{=}1$) and Proto-BagNet (with $\lambda_{L1,s} = 4 \cdot 10^{-2}$, $\lambda_{diss} = 5 \cdot 10^{-3}$, $K = 5$), as well as for dense BagNet and ResNet-50. Our Proto-BagNet performed comparably to the state-of-the-art models (Tab. 1, see confidence intervals in Suppl. Tab. 1), showing that our modifications towards better interpretability did not substantially impair classification performance.

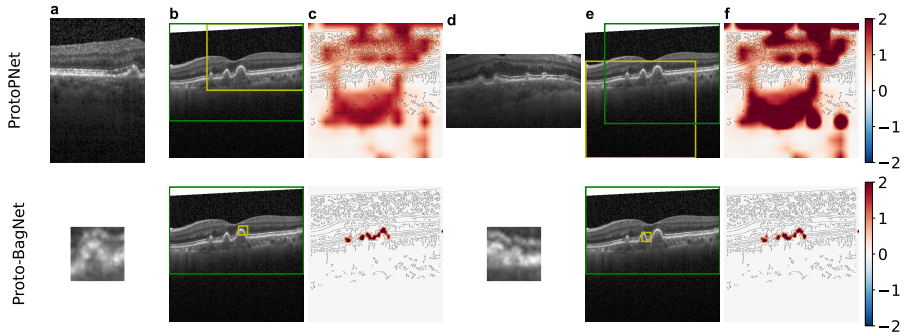

Fig. 2. Example explanations of ProtoPNet and Proto-BagNet. **(a,d)** show two learned prototypes with the highest classification weights. Proto-BagNet's prototypes were magnified for visualization only. **(b,e)** show bounding boxes around regions of highest activation using the visualization technique provided by ProtoPNet (green) and the model's receptive field (yellow). **(c,f)** show prototypes activations of the query image.

3.3 Proto-BagNet provides understandable localized explanations

To qualitatively assess explanations provided by our model, we visualized the two learned prototypes with the highest classification weights for our Proto-BagNet as well as ProtoPNet (Fig. 2a,d). ProtoPNet learned very large prototypes covering almost the entire retina, while Proto-BagNet learned prototypes of small regions of interest with fixed sizes corresponding to its receptive field. For a query image, we then displayed bounding boxes (Fig. 2b,e) around the most similar prototypical part to the learned prototypes. For both models, we first computed the explanation of the prototypical part as the receptive field around the location of the highest prototype similarity (yellow boxes). Due to the large receptive field in ProtoPNet, the explanations were not informative, while the Proto-BagNet yielded small localized patches of the same size as the prototypes themselves. We then computed the explanation as the bounding box around the 95th percentile of the similarity map as in [4], which is not faithful to the model's predictions, as it usually leads to large and similar explanations (green bounding boxes, Fig. 2b,e). In both models, this again led to large bounding boxes around the entire retina, indicating that such prototype explanations may

Table 1. Classification performance for drusen detection on validation and test sets.

	Validation set				Test set			
	Accuracy	AUC	Recall	Precision	Accuracy	AUC	Recall	Precision
ResNet-50	0.991	0.999	0.982	0.986	0.994	0.999	0.992	0.996
dense BagNet	0.990	0.999	0.978	0.985	0.988	0.999	0.976	0.999
ProtoPNet	0.987	0.996	0.975	0.974	0.998	0.999	0.996	0.999
Proto-BagNet	0.978	0.990	0.935	0.981	0.968	0.992	0.940	0.996

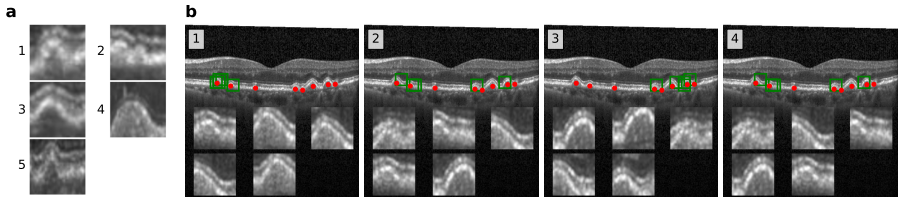

Fig. 3. We show **(a)** the five learned disease prototypes and **(b)** suspicious regions (green boxes, enlarged below) extracted from each prototype similarity map on an example image. Drusen (annotated with red markers) are detected with high precision.

not be useful when evidence may be spread in an image (see similarity maps in Fig. 2c,f), as is often the case in medical images. To summarize, the coarse explanations provided by ProtoPNet were not informative as already highlighted in [7, 19], while the small receptive field of Proto-BagNet leads to fine-grained and localized explanations.

3.4 Proto-BagNet learns meaningful and relevant prototypes

We assessed the clinical meaning and relevance [5, 8] of the prototypes learned by Proto-BagNet by evaluating (1) their interpretability and (2) their coherence as the precision of their corresponding similarity maps at localizing drusen.

We first evaluated the interpretability of the learned prototypes by showing the prototypes without additional context (Fig. 3a) to a clinical expert and asking her if she could understand the concepts encoded in each of them. She described each prototype despite their low resolution as: *(1) soft drusen*; *(2) two drusen in transformation*; *(3) typical drusen*; *(4) drusen with RPE* [7] *thinning and dense substance inside*; and *(5) drusen with RPE thinning probably in transformation*, showing that the learned prototype where semantically meaningful even when seen in isolation, and diverse due to the enforced dissimilarity (Suppl. Fig. 1). Next, we obtained annotations of the 5 training images from which the learned prototypes were extracted (Suppl. Fig. 2). We found that all learned prototypes (i.e. 100%) were extracted from regions labeled as drusen, confirming that Proto-BagNet learns interpretable and meaningful prototypes.

Subsequently, we evaluated the relevance of the learned prototypes on a subset of 40 test images where an ophthalmologist annotated the presence of drusen. On this subset, we calculated the precision of the prototype similarity maps at localizing drusen lesions to assess whether the highlighted prototypical parts contain similar concepts (drusen related) to those encoded by the learned prototypes. We extracted the $k = 5$ prototypical parts as the most discriminative regions similar to each disease prototype (Fig. 3b, more examples in Suppl. Figs. 3, 4), which were also used in the classification mechanism.

[7] Retinal Pigment Epithelium

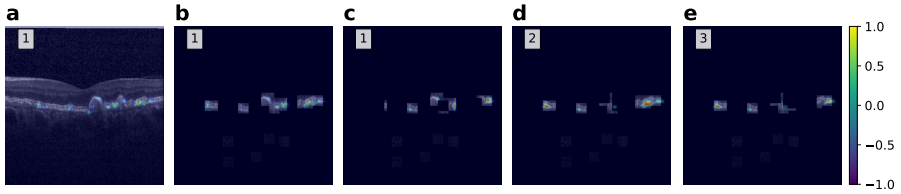

Fig. 4. (a) Example drusen image with overlaid prototype similarity map. **(b)** Occluded image keeping only the top five regions most similar to each prototype. **(c-e)** Images resulting from occluding the most five regions similar to a prototype (1,2,3).

From these, we calculated the precision as the proportion of the prototypical parts (green boxes, magnified on the bottom) which contained annotated drusen lesions (red markers) [13]. The top-k prototypical parts highlighted by the prototype heatmaps contained drusen lesions with high precision ranging from 0.83 ($k = 2$) to 0.87 ($k = 1$) depending on $k \in \{1, \ldots, 5\}$, with $p = 0.84 \pm 0.2$ (mean $\pm$ SD) for $k = 5$.

3.5 Proto-BagNet relies on a faithful decision-making process

Finally, we verified that Proto-BagNet makes decisions based solely on the visual explanation it provides (i.e. on the k input prototypical parts most similar to each prototype). To assess the faithfulness of our model, we applied the model to the test set (example image in Fig. 4a) and then masked all image regions except the top-k prototypical parts identified by the model (Fig. 4b). We reapplied the model to these occluded images and compared the classification output to the output on the original data. The distribution of predicted logits on original images (0.03 ± 0.06, and 0.96 ± 0.15) was similar to that on occluded images (0.03 ± 0.07, and 0.96 ± 0.15), respectively, for healthy and diseased images. The AUC after occlusion was almost similar (0.9918 vs 0.9916) to that obtained without occlusion. We conclude that Proto-BagNet really makes decisions based only on the prototypical parts, and that the explanations provided by the extracted regions are faithful representations of these prototypical parts.

We quantified the importance of each prototype by additionally masking its k prototypical parts in the occluded test images (Fig. 4c-e) and measuring the change in classifier predictions. The predicted drusen probability on healthy images increased by 0.27 ± 0.14 on average when removing healthy prototypical parts (i.e. the region similar to healthy prototypes), while it decreased by 0.42 ± 0.17 when removing disease prototypical parts on drusen images. To summarize, the prototypical parts extracted by the model are indeed evidence for healthy and diseased tissue, respectively, as can be seen by the changes in classifier predictions when occluding them.

4 Discussion and Conclusion

In this work, we proposed Proto-BagNet, an interpretable-by-design prototype-based model that provides faithful and highly localized explanations as well as global interpretability through meaningful prototypes. We evaluated the interpretability of our model through feedback by an ophthalmologist who identified diverse and clinically relevant concepts in the learned prototypes. Furthermore, the model explanations precisely detected drusen lesions in the images. We evaluated the Proto-BagNet for drusen classification on OCT, a solved task in terms of predictive performance (Tab. 1), which lends itself for studying interpretable models due to identifiable and well-characterized regions of interest. However, we noticed that some design choices (e.g., SA layer, k-values) and introduced loss components (dissimilarity and sparsity) enhance interpretability but compete with classification performance (Tab. 1), suggesting that determining the ideal tradeoff might depend on the specific clinical setting. Additionally, the appropriate receptive field size may vary depending on the clinical task and image resolution. In our case, the drusen are small ($< 63\mu m$ [18]) and fit into a patch of size 33×33 and can be changed for other tasks to inject clinical knowledge and adjust for resolution. In a next step, we believe our approach could also be applied to more challenging tasks, and could be useful for the discovery of unknown relevant concepts. In summary, our work may enable prototype-based networks to take a more central stage for realistic task settings, as a promising alternative to post-hoc explanations of black-box models, in particular on medical images.

Acknowledgments. This project was supported by the Hertie Foundation, the German Science Foundation (Excellence Cluster EXC 2064 "Machine Learning—New Perspectives for Science", project number 390727645) and the Carl Zeiss Foundation ("Certification and Foundations of Safe Machine Learning Systems in Healthcare"). PB is a member of the Else Kröner Medical Scientist Kolleg "ClinbrAIn: Artificial Intelligence for Clinical Brain Research". The authors thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting KD.

Disclosure of Interests. The authors declare no competing interests.

References

1. Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., Kim, B.: Sanity checks for saliency maps. Advances in neural information processing systems **31** (2018)
2. Barnett, A.J., Schwartz, F.R., Tao, C., Chen, C., Ren, Y., Lo, J.Y., Rudin, C.: A case-based interpretable deep learning model for classification of mass lesions in digital mammography. Nature Machine Intelligence **3**(12), 1061–1070 (2021)
3. Brendel, W., Bethge, M.: Approximating CNNs with Bag-of-local-Features models works surprisingly well on ImageNet. In: Proc. International Conference on Learning Representations (ICLR) (2019)

4. Chen, C., Li, O., Tao, D., Barnett, A., Rudin, C., Su, J.K.: This looks like that: deep learning for interpretable image recognition. Advances in neural information processing systems **32** (2019)
5. Davoodi, O., Mohammadizadehsamakosh, S., Komeili, M.: On the interpretability of part-prototype based classifiers: a human centric analysis. Scientific Reports **13**(1), 23088 (2023)
6. Fleckenstein, M., Keenan, T.D., Guymer, R.H., Chakravarthy, U., Schmitz-Valckenberg, S., Klaver, C.C., Wong, W.T., Chew, E.Y.: Age-related macular degeneration. Nature reviews Disease primers **7**(1), 31 (2021)
7. Gautam, S., Höhne, M.M.C., Hansen, S., Jenssen, R., Kampffmeyer, M.: This looks more like that: Enhancing self-explaining models by prototypical relevance propagation. Pattern Recognition **136**, 109172 (2023)
8. Ghorbani, A., Wexler, J., Zou, J.Y., Kim, B.: Towards automatic concept-based explanations. Advances in neural information processing systems **32** (2019)
9. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2016)
10. Huang, Q., Xue, M., Huang, W., Zhang, H., Song, J., Jing, Y., Song, M.: Evaluation and improvement of interpretability for self-explainable part-prototype networks. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2011–2020 (2023)
11. Jacovi, A., Goldberg, Y.: Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 4198–4205 (Jul 2020)
12. Kermany, D.S., Goldbaum, M., Cai, W., Valentim, C.C., Liang, H., Baxter, S.L., McKeown, A., Yang, G., Wu, X., Yan, F., et al.: Identifying medical diagnoses and treatable diseases by image-based deep learning. cell **172**(5), 1122–1131 (2018)
13. Kerol, D., Ilanchezian, I., Kühlewein, L., Faber, H., Baumgartner, C.F., Bah, B., Berens, P., Koch, L.M.: Sparse activations for interpretable disease grading. In: Medical Imaging with Deep Learning (2023)
14. Kim, E., Kim, S., Seo, M., Yoon, S.: Xprotonet: diagnosis in chest radiography with global and local explanations. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 15719–15728 (2021)
15. Koh, P.W., Nguyen, T., Tang, Y.S., Mussmann, S., Pierson, E., Kim, B., Liang, P.: Concept bottleneck models. In: International conference on machine learning. pp. 5338–5348. PMLR (2020)
16. Rudin, C.: Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature machine intelligence **1**(5), 206–215 (2019)
17. Selvaraju, R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Grad-Cam, B.: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV). pp. 618–626
18. Silvestri, G., Sillery, E., Henderson, D., Brogan, P., Silvestri, V.: Prevalence of drusen and drusen size in young adults. Investigative Ophthalmology & Visual Science **46**(13), 3298–3298 (2005)
19. Xu-Darme, R., Quénot, G., Chihani, Z., Rousset, M.: Sanity checks for patch visualisation in prototype-based image classification. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023 - Workshops, Vancouver, BC, Canada, June 17-24, 2023. IEEE (2023)

Soft-CAM: Making black box models self-explainable for high-stakes decisions

Kerol Djoumessi, Philipp Berens
Hertie Institute for AI in Brain Health University of Tübingen, Germany
Tübingen AI Center, University of Tübingen, Germany
{kerol.djoumessi-donteu, philipp.berens}@uni-tuebingen.de
https://hertie.ai/

Abstract

Convolutional neural networks (CNNs) are widely used for high-stakes applications like medicine, often surpassing human performance. However, most explanation methods rely on post-hoc attribution, approximating the decision-making process of already trained black-box models. These methods are often sensitive, unreliable, and fail to reflect true model reasoning, limiting their trustworthiness in critical applications. In this work, we introduce SoftCAM, a straightforward yet effective approach that makes standard CNN architectures inherently interpretable. By removing the global average pooling layer and replacing the fully connected classification layer with a convolution-based class evidence layer, SoftCAM preserves spatial information and produces explicit class activation maps that form the basis of the model's predictions. Evaluated on three medical datasets, SoftCAM maintains classification performance while significantly improving both the qualitative and quantitative explanation compared to existing post-hoc methods. Our results demonstrate that CNNs can be inherently interpretable without compromising performance, advancing the development of self-explainable deep learning for high-stakes decision-making.

1 Introduction

Convolutional Neural Networks (CNNs) have revolutionized computer vision by efficiently capturing local patterns, reducing parameters, and accelerating convergence, enabling superior performance in tasks like image recognition and object detection [35, 40]. However, their lack of interpretability limits adoption in high-stakes fields like medicine, where transparency and trust are crucial. To explain CNNs, numerous saliency-based methods have been proposed, including class activation maps (CAM) [61] and their variants [14, 47, 56, 59], gradient-based techniques [48, 50, 51, 53], and even perturbation- or occlusion-based methods [22, 56, 59]. These methods have been widely adopted to explain the decisions of trained black-box models.

Such saliency map-based techniques offer explanations for CNN classifiers that claim to highlight regions in the input image most relevant to the model's prediction. These explanations are generated post-hoc, typically after a model is trained [10, 27]. Studies have shown significant limitations in their effectiveness, especially in clinical settings [5]. Post-hoc saliency methods often lack faithfulness, reliability, and consistency, resulting in explanations that may not accurately reflect the model's decision-making process [3, 58]. Moreover, they struggle to localize relevant regions in medical imaging [4], where the limited availability of ground-truth annotations makes it difficult to assess their trustworthiness. To overcome these challenges, inherently and/or self-explainable models have been introduced [45], designed explicitly to provide interpretable insights by incorporating explanations within their architecture [11, 13, 16, 18, 52]. These models generate more trustworthy

Preprint. Under review.

and faithful explanations that align closely with the model's actual reasoning [28]. However, self-explainable models generally use specific architectures [13, 16, 52], which limits their applicability and generalization to widely used CNN architectures.

Motivated by these challenges, we propose SoftCAM, a straightforward generalization of Class Activation Maps (CAM) that uses a convolution-based classifier to transform any black-box CNN into a self-explainable model. By removing the final pooling layer and replacing the fully connected classification layer with 1x1 convolutions, SoftCAM turns classical CNNs into fully convolutional networks, generating class-specific evidence maps that are directly used for predictions. Our contributions are:

- We introduced SoftCAM, a simple modification to CNNs that enables self-explainability, and experimentally demonstrate that the resulting model preserves classification performance across three clinically relevant medical datasets spanning different imaging modalities.
- We showed that regularizing evidence maps using ElasticNet, a regularizer combining both ridge and lasso penalties, enhances the model's explanations.
- We evaluated five widely used traditional CAM-based post-hoc explanation methods, showing that SoftCAM most often outperforms them across a broad range of explainability metrics and across all three considered medical imaging datasets and modalities.

2 Related work

Deep neural networks (DNNs) are widely used in a variety of fields. However, regulatory frameworks such as the European AI Act require AI-based decisions to be explainable to ensure fairness, transparency, and accountability, allowing users to understand their decision-making process [2, 41]. Explainability can help verify and even improve performance by detecting shortcuts and identifying clinically relevant features, ultimately fostering greater trust in decision-making, especially in fields like healthcare [17].

Existing explainable AI (XAI) methods for image analysis can be broadly categorized into attribution-based and non-attribution-based approaches [10, 27, 57]. Attribution-based methods explain "where" important features exist in the input by generating saliency or heatmaps that assign importance scores to individual pixels or regions, helping visualize their contribution to the model's decision. These include perturbation-based methods [29] and class activation maps [26], which consist of both gradient-based [47, 51, 53] and gradient-free approaches [44, 56, 61]. In contrast, non-attribution-based approaches explain "why" a decision was made without relying on importance scores, instead using techniques such as concept-based (ACE [23], TCAV [32], CBM [33]), prototype-based [15, 16], or counterfactual-based [7, 12, 24, 52] methods to analyze model behavior from different perspectives. These approaches also differ in how explanations are obtained and which architectures they can be applied to. Attribution-based methods typically provide post-hoc, local explanations by offering input-specific insights into black-box CNN models after training. In contrast, non-attribution-based methods are generally inherently interpretable by design, promoting transparency and enabling global understanding of the model's decision-making process across the entire dataset [45]. However, in some cases, self-explainable models may be less effective for complex tasks, highlighting a tradeoff between interpretability and performance, where increasing transparency may sometimes come at the cost of classification performance [57].

Our method, SoftCAM, relies on an explicit class-evidence layer based on convolutional operations for classification, offering class-specific, attribution-based explanations while remaining inherently interpretable, unlike post-hoc approaches. Moreover, it maintains predictive performance comparable to its corresponding non-interpretable black-box models. Closely related work includes [6] and [18]. In [6], a dual-branch approach is used, where one branch leverages a traditional black-box CNN model (ResNet-18) for classification, and the second branch uses weight-sharing for post-hoc explanations, requiring two forward passes for inference. In the second branch for post-hoc explanation, the global average pooling layer (GAP) is removed, and the linear classifier is replaced by convolutional layers that share weights during inference to generate class-specific activation maps. In contrast, SoftCAM is trained end-to-end, providing both predictions and explanations in a single forward pass, eliminating the need for additional computational overhead or weight sharing. Furthermore, while [18] uses explicit class-evidence maps to enhance the explanation of a self-explainable bag-of-local-feature model (BagNet [13]), our method transforms black-box models into self-explainable models. We

evaluated our approach on a range of medical datasets, comparing the resulting explanations to various post-hoc attribution-based methods, including both gradient-free and gradient-based techniques.

3 Method

Preliminaries Given an input image $\mathbf{X} \in \mathbb{R}^{H_X \times W_X \times C_X}$ with height H_X, width W_X, and the number of channels C_X, consider a CNN network f_θ that maps $\mathbf{X}$ to a probability distribution $\hat{\mathbf{y}} = f_\theta(\mathbf{X}) \in \mathbb{R}^C$, where C is the number of classes, and $y^c \in \mathbf{y}$ represents the predicted probability for class c. The network consists of a feature extractor g_ϕ, and a classifier layer h_ψ, with learnable parameters ϕ and ψ. The feature extractor generates a feature map $\mathbf{Z} = g_\phi(\mathbf{X}) \in \mathbb{R}^{N \times M \times D}$, where $N \times M$ denotes the spatial size and D is the feature dimension (e.g., $D = 2048$ for most ResNet variants). The classifier then predicts the final output based on the extracted features. Let $\mathcal{A} = \{\mathbf{A}_k\}_{k=1}^D$, the set of activation maps obtained from the feature extractor, where A_k is the activation of the k-th neuron. Let, $S^c_{\text{Map}} \in \mathbb{R}^{N \times M}$ be the 2D saliency map, providing a visual explanation of the model's prediction for class c. This paper explores how to train self-explainable CNNs to simultaneously generate both the prediction y^c and its corresponding explanation S^c_{Map}.

Traditional CNN architectures employ a GAP layer to reduce the feature map to $1 \times D$, followed by a classification module consisting of one or more linear fully connected layers (FCLs) to generate the final prediction. Post-hoc methods are then typically used to explain the model's decision.

3.1 CAM-based methods

Class Activation Maps (CAMs) [61] are closely related to our approach, offering local visual explanations of CNN predictions by generating saliency maps for individual inputs. CAM achieves this by linearly combining the feature maps from the final convolutional layers with importance coefficients from the FCL classifier, thereby producing class-wise attribution maps as follows:

$$S^c_{\text{CAM}}(x_1, x_2) = \sum_{k=1}^{D} w^c_k A_k(x_1, x_2), \tag{1}$$

where $A_k(x_1, x_2)$ is the activation of neuron k in the feature map at spatial location (x_1, x_2), and w^c_k denotes the importance weight associated with class c for unit k in the fully connected layer.

Originally, CAM was designed for CNNs with GAP and FCL, but has been extended to gradient-based methods using class score gradients to compute importance weights [14, 47, 48]. This extension enabled CAM-based techniques to be applied to a broader range of CNN architectures, particularly those where the GAP layer is followed by multiple FCLs, as seen in models like VGG [49] and InceptionV3 [54]. For example, GradCAM [47] extends the original CAM approach by backpropagating the gradient from a target class to the input layer to highlight the image regions that strongly influence the model's prediction. GradCAM is formulated as

$$S^c_{\text{Grad-CAM}}(x_1, x_2) = \text{ReLU}\left(\sum_{k=1}^{D} w^c_k A_k(x_1, x_2)\right), \tag{2}$$

where the weight coefficients are computed as $w^c_k = \frac{1}{N \times M} \sum_i^N \sum_j^N \frac{\partial y^c}{\partial A_k(i,j)}$. Here, $A_k(i, j)$ is the activation value at location (i, j) on A_k, and the rectified linear unit (ReLU) is applied to ensure that the final activation map considers only the features that positively influence class c. Following GradCAM, several variations have been proposed, including gradient-based approaches such as SmoothGrad [50], GradCAM++ [14], guided-backpropagation [51], and integrated gradients [53], as well as gradient-free methods like ScoreCAM [56], LayerCAM [30], and OptiCAM [59]. Gradient-based methods primarily differ in how gradients are aggregated to compute importance weights, while gradient-free methods mainly vary in how the weights are computed.

Despite the success of class activation map-based methods in explaining CNN classifiers, including medical applications [8], they have a key limitation: they rely on already trained models and provide post-hoc explanations, which may not accurately reflect the model's true decision-making process. Additionally, gradient-based methods face inherent challenges such as gradient saturation, where DNN gradients tend to diminish, and false confidence, where the highest activation map weight does

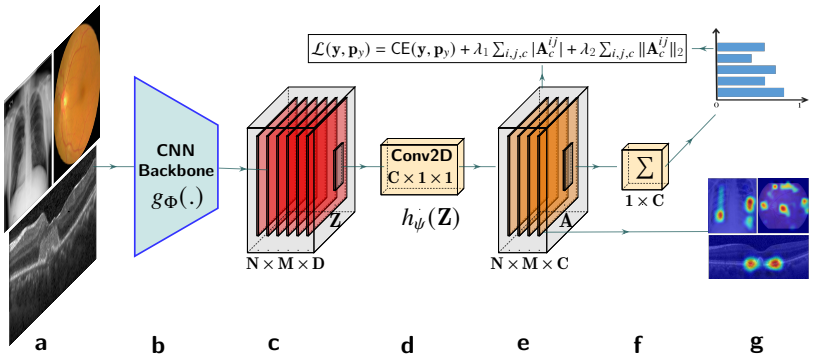

Figure 1: **Overview of softCAM for making black-box CNNs inherently interpretable.** (**a**) Input image. (**b**) The CNN backbone consists of all layers before the global average pooling layer. (**c**) Feature map generated by the backbone. (**d**) Classifier module with C convolutional kernels of size 1×1. (**e**) Self-explainable class activation maps $\mathbf{A}$, obtained from the classifier with ElasticNet penalty applied to it to enhance interpretability. (**f**) Final predictions are derived directly from the evidence maps via spatial average pooling followed by the softmax function. Class-specific evidence maps (**g**) are upsampled and overlaid on the input to visualize the model's decision-making process.

not necessarily correspond to the greatest increase in confidence [56]. On the other hand, gradient-free methods are computationally and memory-intensive, as they often require multiple forward passes on perturbed inputs. Finally, CAM-based methods are easy to implement in CNNs with clearly defined spatial feature maps, but face challenges in multi-branch architectures like InceptionV3 due to the complexity of integrating diverse feature maps from parallel convolutional paths.

3.2 Improving CAMs for self-explanability

Motivated by the limitations of post-hoc class activation map-based methods in interpreting CNN classifiers, we introduce SoftCAM (Fig. 1), a straightforward modification of black-box CNN classifiers that makes them self-explainable and inherently interpretable. SoftCAM achieves this by replacing the fully connected classification layer in classical CNNs with an explicit class-evidence convolutional layer, preserving spatial information and providing explanations in a single forward pass, eliminating the need and computational overhead for post-hoc techniques.

We make black-box CNN architectures self-explainable by modifying how predictions are obtained. Any FCL of size $b_1 \times b_2$, where b_1 and b_2 denote the number of input and output features, respectively, can be equivalently expressed as a 1×1 convolutional layer with b_1 input channels and b_2 output channels [18]. This allows us to replace FCL classifiers in standard CNN architectures with convolutions, removing the GAP layer before classification, while preserving model complexity and spatial localization. The new classifier module h consists of convolutional layers (Fig. 1d) with C convolution kernels of size 1×1 and unit stride, producing class evidence maps (Fig. 1e)

$$\mathbf{A} = h_\psi(\mathbf{Z}) \in \mathbb{R}^{M \times N \times C}, \tag{3}$$

where ψ is a learnable parameter. Indeed, h_ψ can be viewed as an explainable, soft generalization of classical post-hoc attribution methods (Eq. 1, 2), mapping the low-dimensional feature volume $\mathbf{Z}$ into an interpretable, class-wise activation volume $\mathbf{A}$ whose reduced channel dimension corresponds to the number of target classes. Unlike CAM (Eq. 1) and GradCAM (Eq. 2), which generate post-hoc heuristic explanations, our approach leverages the feature map volume from the backbone and applies a parameterized function h_ψ to directly produce class activation maps that are used for prediction. In contrast to classical CAM-based methods, the importance weights are not explicitly defined but are implicitly learned and encoded within the classifier's parameters.

The resulting architecture is a fully convolutional, self-explainable model, where the final predicted probabilities are computed from the evidence map (Fig.1e), without introducing additional parameters:

$$\hat{\mathbf{y}} = \mathrm{Softmax}\Big(\mathrm{AvgPool}\Big(h_\psi\big(g_\Phi(\mathbf{X})\big)\Big)\Big) \in \mathbb{R}^{1 \times C}. \tag{4}$$

Additionally, the class evidence maps $\mathbf{A}$ serve as built-in explanations, directly representing the contribution of individual input regions to the final prediction (Fig. 1g). Replacing linear FCL layers with convolutional operations offers several advantages. Due to the shift-invariance and position-agnostic properties of CNNs, all image regions are weighted equally when forming the final classification (Fig. 1f). As a result, input feature patches with high activations in the evidence maps contribute most significantly and linearly to the prediction. This behavior mirrors that of simple linear models, where each value in the activation map has a direct and interpretable impact on the output.

3.3 Regularizing SoftCAM for interpretability

By using explicit class-evidence maps, the model can be trained directly with regularization applied to these maps to enhance interpretability. In practice, we apply an ElasticNet regularization [62], which linearly combines the ℓ_1 (Lasso) and ℓ_2 (Ridge) penalties, leading to the following loss function:

$$\mathcal{L}(\mathbf{y}, \hat{\mathbf{y}}) = \text{CE}(\mathbf{y}, \hat{\mathbf{y}}) + \lambda_1 \sum_{i,j,c} |\mathbf{A}_c^{ij}| + \lambda_2 \sum_{i,j,c} ||\mathbf{A}_c^{ij}||_2. \tag{5}$$

Here, CE denotes the cross-entropy loss, and $\mathbf{y}$ represents the reference class labels. When $\lambda_2 = 0$, the ElasticNet penalty becomes the Lasso penalty, which promotes sparsity in the class evidence maps [18] by removing less informative activations, making it particularly useful for tasks where precision in explanations is crucial. In contrast, when $\lambda_1 = 0$, the ElasticNet reduces to the Ridge penalty, which reduces irrelevant activations without forcing them to zero, which is beneficial when minimizing false negatives is a priority. ElasticNet thus provides a balance between Lasso and Ridge penalties, balancing sparsity and smoothness in the resulting activation maps.

Visualizing explanations. The evidence map generated by SoftCAM is upsampled to the input resolution for visualization (Fig. 1g). Like most CAM-based methods, such as GradCAM [47], ScoreCAM [56], and LayerCAM [30] that operate on the final convolutional layer, SoftCAM's explanations are limited by the resolution of the backbone (e.g., 16×16 for VGG-16/ResNet-50 with 512×512 input) due to pooling and striding, leading to lower-resolution saliency maps. However, by introducing the class evidence and classification layer directly atop features, SoftCAM regularizes the evidence map, making it less coarse and thereby enhancing localization. In contrast, gradient-based methods like Integrated Gradients [53] and Guided Backpropagation [51] produce high-resolution saliency maps by computing pixel-level gradients, which may lead to noisy maps, especially when the region of interest spans a broader area, as commonly observed in Chest X-ray images.

Comparison with other approaches. Unlike post-hoc attribution-based approaches, our method is inherently interpretable from the classification layer and maintains performance comparable to its black-box counterpart, without a significant trade-off, even when regularization is applied to enhance explainability. Compared to [18], our method extends from the concept of interpretable bag-of-local models to general black-box CNN architectures and generalizes the regularization from Lasso to ElasticNet, with extensive evaluations across multiple datasets using a broad range of explanability metrics. Compared to [6], our method is trained end-to-end and does not require post-hoc processing, weight sharing between branches, or an additional forward pass to generate explanations.

4 Experimental setup

Datasets. We evaluated our approach on three publicly available medical datasets spanning three imaging modalities: the Kaggle Diabetic Retinopathy (DR) [20], Retinal OCT [31], and the RSNA Chest X-Ray (CXR) [1]. The first dataset comprised high-resolution retinal color fundus images, each labeled with a DR severity score ranging from 0 (No DR) to 4 (Proliferative DR). The second dataset included retinal OCT B-scans images categorized into Drusen, Diabetic macular edema, Choroidal neovascularization, and Normal cases. The final dataset consisted of high-resolution frontal-view chest radiographs labeled for pneumonia detection, with bounding box annotations for pneumonia cases. Additionally, lesion annotations were obtained for 65 DR images from the Kaggle dataset [17] and 40 Drusen images from the retinal OCT dataset [16]. Each dataset was split into training, validation, and test sets using different dataset-specific train-validation-test proportions, ensuring that all samples from the same patient were assigned to the same split to prevent data leakage. For full details, see Appendix A.1.

Table 1: Classification performance for binary disease detection on the test sets. We denote the SoftCAM versions of ResNet and VGG with a *.

	Kaggle Fundus				OCT retinal				RSNA CXR	
	Binary		Multi-class		Binary		Multi-class		Binary	
	Acc.	AUC	Acc.	κ	Acc.	AUC	Acc.	κ	Acc.	AUC
VGG-16	0.907	0.938	0.863	0.835	0.994	1.0	0.967	0.955	0.952	0.989
dense VGG*	0.915	0.942	0.861	0.834	0.994	1.0	0.963	0.947	0.957	0.999
sparse VGG*	0.911	0.938	0.859	0.827	0.988	0.999	0.947	0.929	0.953	0.990
ResNet-50	0.899	0.923	0.850	0.800	0.994	0.999	0.970	0.963	0.953	0.988
dense ResNet*	0.899	0.926	0.851	0.811	0.994	1.0	0.974	0.960	0.942	0.983
sparse ResNet*	0.895	0.923	0.851	0.801	0.996	1.0	0.963	0.955	0.941	0.979

Baseline models. The effectiveness of our method was evaluated using two widely used black-box CNN architectures: ResNet-50 [25] and VGG-16 [49]. They differ primarily in the design of their classification heads, where ResNet employs a single fully connected layer, while VGG uses multiple. In both models, we explicitly replaced the classification head with our convolutional evidence map layer, adapting the architecture to enable interpretability (see Appendix A.2). The models were sourced from Torchvision [36], initialized with pre-trained weights from ImageNet, and fine-tuned using a consistent setup[1]. Training was performed over 70 epochs with a mini-batch size of 16 on an NVIDIA A40 GPU using PyTorch [42]. A range of data augmentation and preprocessing techniques was applied (see Appendix A.3). For complete training details, see Appendix A.4.

Baseline CAM-based methods. We qualitatively and quantitatively assessed the explanations generated by our method (SoftCAM) against post-hoc explanation techniques from several state-of-the-art class attribution map-based methods, applied to their respective black-box models. Specifically, we compared our approach with gradient-based methods, including GradCAM [47], Integrated Gradient (Itgd Grad.) [53], Guided Backpropagation (Guided BP) [51], as well as gradient-free methods such as ScoreCAM [56] and LayerCAM ([30]). Guided BP and Itgd Grad. have consistently performed well in producing saliency maps for explaining black-box CNN classifiers on retinal images [8, 17], while GradCAM has shown strong localization performance for chest X-ray interpretation [46]. Gradient-based methods were implemented from Captum [34], whereas gradient-free methods were implemented via TorchCAM [21]. For full descriptions of these methods, see Appendix A.5.

Evaluation metrics. Models were evaluated on both classification performance and explainability. For binary tasks, performance was measured using accuracy and AUC, while for multi-class tasks, accuracy and the quadratic Cohen's kappa score were used. AUC reflects class separability, whereas the kappa score captures agreement beyond chance. To assess explainability, we employed several quantitative metrics: Top-k localization precision [18], activation precision [9, 43], activation consistency [18], and faithfulness [16, 39]. We further extended activation precision to define activation sensitivity. For full descriptions of the explainability metrics, see Appendix B.

5 Results

5.1 Making black box CNNs self-explainable maintains classification performance

We first evaluated our method on clinically relevant classification tasks, including retinal disease classification from color fundus and OCT retinal images, as well as pneumonia detection from chest X-rays. For the fundus and OCT retinal datasets, both binary classification ({0} vs. {1-4} for fundus and Normal vs. Drusen for OCT) and multi-class classification tasks were considered, as reference labels were available. In contrast, the RSNA CXR dataset only included labels for pneumonia detection, restricting the evaluation to the binary task. For each CNN architecture, the "dense" model corresponds to our method without regularization ($\lambda_1 = \lambda_2 = 0$), while the "sparse" model is obtained by applying a lasso penalty ($\lambda_2 = 0$) and choosing an appropriate value for λ_1 (e.g. $\lambda_1 = 1.10^{-6}$ for VGG and $\lambda_1 = 5.10^{-5}$ for ResNet on the fundus dataset). The sparsity parameter was selected by balancing classification accuracy and AUC on the validation set (see Appendix C).

[1] Our code with datasets is available at `https://anonymous.4open.science/r/SoftCAM-E1A3/`

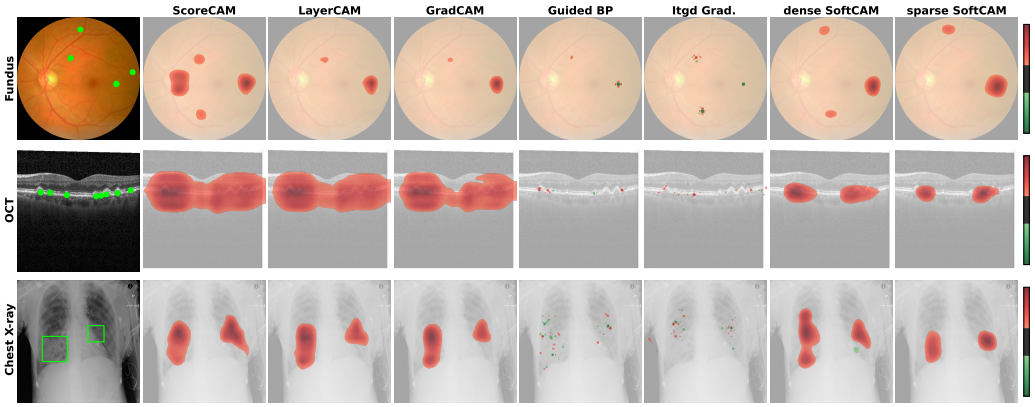

Figure 2: **Example explanations generated by different methods from ResNet-50**. The first column shows disease images with reference annotations, indicated by green markers or bounding boxes. Each row, from top to bottom, corresponds to fundus, OCT, and Chest X-ray images, respectively. The next five columns present saliency maps generated by post-hoc explanation methods, gradient-free (ScoreCAM, LayerCAM) and gradient-based (GradCAM, Guided BP, Itgd Grad). The final two columns showcase our proposed inherently interpretable dense and sparse SoftCAM explanations.

Our results show that SoftCAM models, which use explicit self-explainable class evidence maps, preserve classification performance comparable to their corresponding black-box counterparts (Tab. 1). Moreover, introducing the Lasso regularization penalty on the class evidence map did not significantly degrade performance; in some cases, it even led to slight improvement. These findings suggest that using convolutional layers in the classification head is an effective and promising approach for developing high-performing, self-explainable CNN models.

5.2 SoftCAM provides inherently interpretable visual explanations

We qualitatively compared the evidence maps of SoftCAM variants with saliency maps generated by the five state-of-the-art CAM-based methods. Overall, our method produced more visually interpretable maps with high evidence regions centered on annotated lesions (Fig. 2). We observed that the regions highlighted by the sparse SoftCAM models are mostly a subset of those identified by the dense SoftCAM, reflecting the effect of the sparsity constraint. Additional results, including those for VGG-16 and other illustrative examples, are provided in Appendix D.1.

On healthy images, sparse SoftCAM evidence maps exhibited overall more negative activations, in contrast to the positive activations observed on disease images. To assess this quantitatively, we computed the activation consistency [18], calculating the proportion of positive and negative activations across disease and healthy samples. These findings were consistent with the qualitative observations (e.g. dense vs. sparse SoftCAM on the fundus dataset using ResNet: 0.55 vs. 0.27 for the proportion of positive activation on disease images). For full analysis, see Appendix D.2.

5.3 SoftCAM provides localized and faithful explanations

To quantitatively assess the explanations provided by our SoftCAM evidence maps in comparison to post-hoc saliency methods, we first evaluated their localization precision, which measures how well the highlighted regions in the explanation maps align with human-annotated ground truth. Following [18], we computed the Top-k (k=30) localization precision by upsampling each explanation map to the input resolution, splitting it into non-overlapping 33x33 patches, and calculating the proportion of positively activated patches that overlap with ground truth annotations. Despite being inherently interpretable, SoftCAM explanations performed competitively overall in terms of localization precision (Fig. 3; Appendix D.3). Notably, the sparse SoftCAM with the ResNet backbone outperformed all other methods with the highest top-k precision (see Appendix D.3, D.4), and ranked second only in top-3 precision on the fundus dataset (Fig. 3), behind Guided BP, which benefits from high-resolution saliency maps. Furthermore, we observed that SoftCAM typically achieved higher

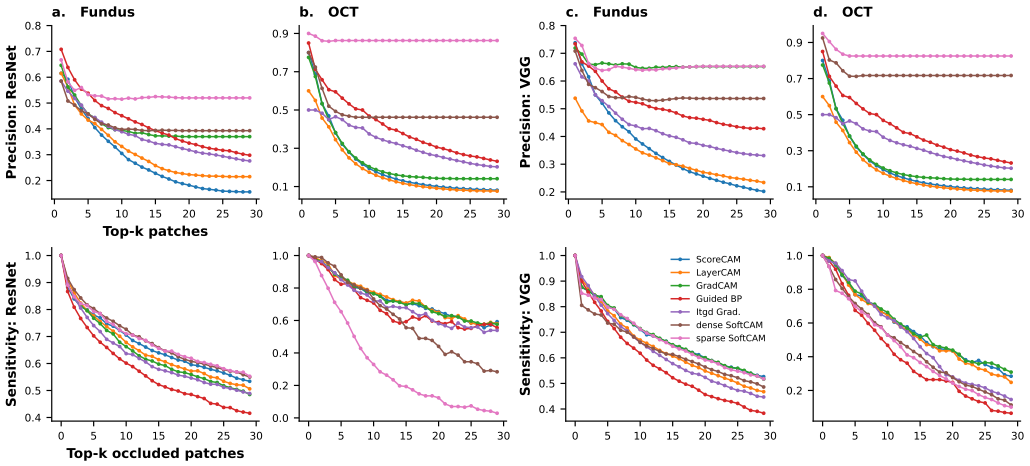

Figure 3: **Quantitative evaluation of explanations generated by different methods**. The first row shows the localization precision of the saliency maps on the Fundus and OCT datasets, evaluated against their respective ground truth. The second row presents the sensitivity analysis assessing the faithfulness of the generated explanations. Columns **a,b** show ResNet results, while **c,d** correspond to VGG. Higher precision means better localization; lower sensitivity implies more reliable explanations.

precision with fewer top-K regions, particularly on the Fundus and OCT datasets. This suggests that SoftCAM more consistently highlighted fewer, yet truly relevant regions, whereas post-hoc methods tended to produce broader and less specific activations, resulting in higher false-positive rates.

Subsequently, we evaluated the faithfulness (also referred to as sensitivity) of the evidence maps generated by our SoftCAM approach, in comparison to post-hoc saliency maps. Sensitivity analysis evaluates how much the highly activated regions in an explanation map contribute to the model's prediction [39], thereby assessing whether the highlighted areas actually influence the model's decision-making process. To do this, we split the input images into non-overlapping 33x33 patches, then progressively removed the top-ranked patches (based on attribution scores) and observed the relative change in model confidence. We conducted this evaluation on samples that were correctly predicted by both the black-box CNNs and their corresponding dense and sparse SoftCAM variants in the test sets. We found that the sparse SoftCAM generally outperformed other methods, notably on the OCT and RSNA datasets (Fig. 3; Appendix D.3, D.4). On the fundus dataset, both the dense and sparse SoftCAM models performed slightly below the best-performing post-hoc methods, with Guided BP yielding the highest sensitivity scores, followed by Integrated Gradients (Fig. 3). On the OCT dataset, sparse and dense SoftCAM outperformed all post-hoc methods when using the ResNet model and ranked second and third, respectively, with the VGG model. Finally, on the RSNA dataset, sparse SoftCAM achieved the best sensitivity scores, outperforming all other methods, while dense SoftCAM ranked second with ResNet and third with VGG (see Appendix D.4).

5.4 Ridge regularization improves explanation for large disease regions

Since the CXR dataset provided larger bounding boxes localizing disease regions, unlike the point-wise lesion annotations available in the fundus and OCT datasets, we computed activation precision [9, 43], which measures the proportion of the class-guided explanation that fall within the ground-truth bounding boxes, emphasizing precision by penalizing only false positives. However, it does not account for sensitivity or penalize false negatives. To address this limitation, we extended this metric to activation sensitivity (see Appendix B.2), which penalizes false negatives to better assess the explanation completeness, especially important in clinical imaging tasks where missing relevant regions can be critical, such as in multi-focal infectious diseases like pneumonia like pneumonia [37]. We further investigated how different regularization strategies affect explanation quality. While Lasso regularization promoted sparsity by shrinking some activations to zeros, ridge regularization encouraged small (but nonzero) values, resulting in denser evidence maps. To evaluate this, we trained

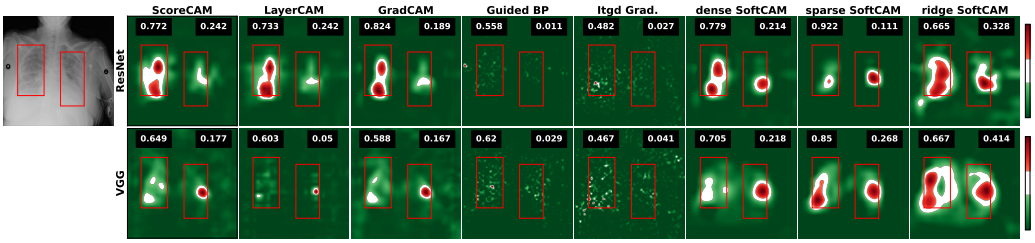

Figure 4: **Example of localization evaluation on the CXR dataset for pneumonia detection**. The first row shows saliency maps generated by different methods from the ResNet model, and the second row from the VGG model. Ground-truth bounding boxes are overlaid on each map, with the top-right value indicating the activation precision, while the top-left value indicates the activation sensitivity.

a ridge SoftCAM model ($\lambda_1 = 0$) and compared its performance to dense and sparse SoftCAM, as well as to the post-hoc explanation methods. The ridge penalty values were selected to balance classification performance ($\lambda_2 = 7.10^{-5}$ vs. $\lambda_2 = 2.10^{-4}$ for ResNet and VGG; see Appendix E.1).

Under comparable classification performance (Acc.=0.95 for Ridge ResNet*, and VGG*), we found that all SoftCAM variants (dense, sparse, and ridge) generally outperformed the evaluated posthoc methods in both activation precision and activation sensitivity (Fig. 4; Appendix E.2, E.3). Specifically, sparse SoftCAM achieved the highest activation precision, while ridge SoftCAM excelled in activation sensitivity. Dense SoftCAM consistently performed in between, underscoring the importance of balancing lasso and ridge regularization via ElasticNet to adapt to varying interpretability needs.

5.5 SoftCAM provides resource-efficient and faithful explanations for multi-class tasks

Finally, we extended our method to the multi-class setting for retinal disease diagnosis. We retrained the same training setup as for the binary tasks, adjusting the output classes in the evidence layer to 5 for DR grading (fundus dataset) and 4 for retinal disease classification (OCT dataset). Given the small size of retinal lesions, we used Lasso regularization, selecting λ_1 values that balanced performance (e.g. $\lambda_1 = 9.10^{-4}$ vs. $\lambda_1 = 3.10^{-6}$ for ResNet and VGG on the OCT dataset; Appendix F.1). Both dense and sparse models achieved performance comparable to their respective black-box baselines (Tab.1), with a slight improvement in Kappa on the fundus dataset when using the ResNet backbone.

As no ground-truth lesion annotations were available for the multi-class tasks, we evaluated the faithfulness of the explanations by measuring their contribution to model predictions. For correctly classified test samples, we progressively removed top-k ($k = 30$) ranked patches (based on the explanation maps; see Sec.5.3) and tracked the average drop in class confidence. In both tasks, the dense and sparse SoftCAM achieved superior performance, with sparse SoftCAM yielding the lowest area under the deletion curve, indicating the highest faithfulness (see Appendix F.2, F.3).

Notably, the sparse SoftCAM produced class-wise explanations that aligned well with class model confidence, showing minimal evidence in healthy classes (Fig.5; Appendix F.4, F.5 for VGG and more examples). In the case of DR detection, a progressive disease, it is expected that images labeled with grade x, where $1 < x < 5$, may still exhibit features from earlier stages, consistent with explanations. Unlike post-hoc CAM-based methods, which require backpropagation or perturbation for each class, SoftCAM generates class-specific explanations during prediction in a single forward pass, making it more resource-efficient.

6 Discussion

Here, we introduced SoftCAM, a straightforward yet effective approach for transforming black-box CNN models into inherently interpretable architectures. We tested SoftCAM on a diverse range of medical imaging tasks, including color fundus photographs, retinal OCT scans, and Chest X-rays for disease diagnosis. Importantly, SoftCAM-variants maintained performance comparable to that of the original CNN models for classification and generally outperformed post-hoc explainability techniques. SoftCAM produces explicit class evidence maps that directly contribute to the model's

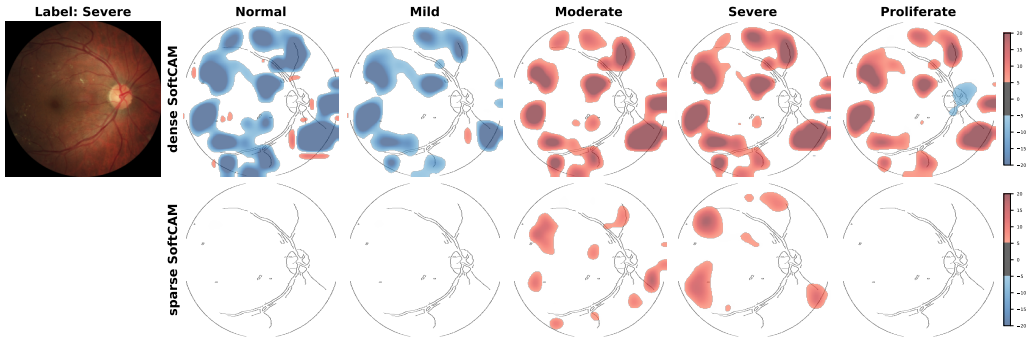

Figure 5: **Examples of multi-class explanations using ResNet**. For a severe DR example from the Kaggle dataset, the first row shows class-specific dense SoftCAM evidence map explanations, while the second presents explanations from the sparse SoftCAM.

prediction. This integration enables single forward-pass generation of explanations aligned with the classification output, resulting in resource-efficient and self-explainable CNNs.

We evaluated our method for two widely used CNN backbones: ResNet-50 and VGG-16, assessing both classification performance and explainability. Despite some differences, both ResNet and VGG models employ large receptive fields, resulting in low-resolution feature maps. Consequently, the class-evidence layer operates on coarse feature maps, producing coarse-grained explanations. In the future, we could explore the integration of SoftCAM with other standard architectures like ViT [19].

Our work presents a major step forward in the development of powerful self-explainable models, demonstrating that interpretable-by-design architectures can preserve, and in some cases even improve upon, the classification performance of state-of-the-art models by modifying standard, well-tested CNN architectures without the need for complicated additional concepts such as prototypes [15]. Beyond performance, SoftCAM provides deeper insights into the model-decision-making process, offering a powerful tool for understanding mistakes and detecting spurious correlations, without relying on widely used post-hoc explanation methods. By leveraging ElasticNet regularization, which is task-specific, user can flexibly balance localization precision and sensitivity according to their application needs. This is especially relevant for CNN-based classifiers deployed in high-stakes decision-making contexts. We hope this contribution will pave the way toward designing more accurate and interpretable CNN models, ultimately fostering trust, adoption, and integration in critical real-world settings such as in medicine.

Acknowledgments and Disclosure of Funding

This project was supported by the Hertie Foundation, the German Science Foundation (Excellence Cluster EXC 2064 “Machine Learning—New Perspectives for Science”, project number 390727645). The authors thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting KD.

References

[1] Rsna pneumonia detection challenge, 2018.

[2] EU Artificial Intelligence Act. The eu artificial intelligence act, 2024.

[3] 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.

[4] José Pereira Amorim, Pedro Henriques Abreu, João Santos, and Henning Müller. Evaluating post-hoc interpretability with intrinsic interpretability. *arXiv preprint arXiv:2305.03002*, 2023.

[5] N Arun, N Gaw, P Singh, K Chang, M Aggarwal, B Chen, et al. Assessing the (un) trustworthiness of saliency maps for localizing abnormalities in medical imaging. arxiv. *arXiv preprint arXiv:2008.02766*, 2020.

[6] Marc Aubreville, Miguel Goncalves, Christian Knipfer, Nicolai Oetter, Tobias Würfl, Helmut Neumann, Florian Stelzle, Christopher Bohr, and Andreas Maier. Transferability of deep learning algorithms for malignancy detection in confocal laser endomicroscopy images from different anatomical locations of the upper gastrointestinal tract. In *Biomedical Engineering Systems and Technologies: 11th International Joint Conference, BIOSTEC 2018, Funchal, Madeira, Portugal, January 19–21, 2018, Revised Selected Papers 11*, pages 67–85. Springer, 2019.

[7] Maximilian Augustin, Valentyn Boreiko, Francesco Croce, and Matthias Hein. Diffusion visual counterfactual explanations. *Advances in Neural Information Processing Systems*, 35:364–377, 2022.

[8] Murat Seçkin Ayhan, Louis Benedikt Kümmerle, Laura Kühlewein, Werner Inhoffen, Gulnar Aliyeva, Focke Ziemssen, and Philipp Berens. Clinical validation of saliency maps for understanding deep neural networks in ophthalmology. *Medical Image Analysis*, 77:102364, 2022.

[9] Alina Jade Barnett, Fides Regina Schwartz, Chaofan Tao, Chaofan Chen, Yinhao Ren, Joseph Y Lo, and Cynthia Rudin. A case-based interpretable deep learning model for classification of mass lesions in digital mammography. *Nature Machine Intelligence*, 3(12):1061–1070, 2021.

[10] Subrato Bharati, M Rubaiyat Hossain Mondal, and Prajoy Podder. A review on explainable artificial intelligence for healthcare: why, how, and when? *IEEE Transactions on Artificial Intelligence*, 2023.

[11] Moritz Böhle, Mario Fritz, and Bernt Schiele. B-cos networks: Alignment is all we need for interpretability. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10329–10338, 2022.

[12] Valentyn Boreiko, Indu Ilanchezian, Murat Seçkin Ayhan, Sarah Müller, Lisa M Koch, Hanna Faber, Philipp Berens, and Matthias Hein. Visual explanations for the detection of diabetic retinopathy from retinal fundus images. In *International conference on medical image computing and computer-assisted intervention*, pages 539–549. Springer, 2022.

[13] Wieland Brendel and Matthias Bethge. Approximating cnns with bag-of-local-features models works surprisingly well on imagenet. *International Conference on Learning Representations*, 2019.

[14] Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE winter conference on applications of computer vision (WACV)*, pages 839–847. IEEE, 2018.

[15] 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.

[16] Kerol Djoumessi, Bubacarr Bah, Laura Kühlewein, Philipp Berens, and Lisa Koch. This actually looks like that: Proto-bagnets for local and global interpretability-by-design. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 718–728, 2024.

[17] Kerol Djoumessi, Ziwei Huang, Laura Kuehlewein, Annekatrin Rickmann, Natalia Simon, Lisa M Koch, and Philipp Berens. An inherently interpretable ai model improves screening speed and accuracy for early diabetic retinopathy. *medRxiv*, pages 2024–06, 2024.

[18] Kerol R Djoumessi Donteu, Indu Ilanchezian, Laura Kühlewein, Hanna Faber, Christian F Baumgartner, Bubacarr Bah, Philipp Berens, and Lisa M Koch. Sparse activations for interpretable disease grading. In *Medical Imaging with Deep Learning*, 2023.

[19] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*, 2021.

[20] Emma Dugas, Jorge Jared, and Will Cukierski. Diabetic retinopathy detection, 2015.

[21] François-Guillaume Fernandez. Torchcam: class activation explorer. `https://github.com/frgfm/torch-cam`, March 2020.

[22] Ruth Fong, Mandela Patrick, and Andrea Vedaldi. Understanding deep networks via extremal perturbations and smooth masks. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2950–2958, 2019.

[23] Amirata Ghorbani, James Wexler, James Y Zou, and Been Kim. Towards automatic concept-based explanations. *Advances in neural information processing systems*, 32, 2019.

[24] Riccardo Guidotti. Counterfactual explanations and how to find them: literature review and benchmarking. *Data Mining and Knowledge Discovery*, 38(5):2770–2824, 2024.

[25] 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.

[26] Mingwei He, Bohan Li, and Songlin Sun. A survey of class activation mapping for the interpretability of convolution neural networks. In *International Conference On Signal And Information Processing, Networking And Computers*, pages 399–407. Springer, 2022.

[27] Md Imran Hossain, Ghada Zamzmi, Peter R Mouton, Md Sirajus Salekin, Yu Sun, and Dmitry Goldgof. Explainable ai for medical data: Current methods, limitations, and future directions. *ACM Computing Surveys*, 2023.

[28] Junlin Hou, Sicen Liu, Yequan Bie, Hongmei Wang, Andong Tan, Luyang Luo, and Hao Chen. Self-explainable ai for medical image analysis: A survey and new outlooks. *arXiv preprint arXiv:2410.02331*, 2024.

[29] Maksims Ivanovs, Roberts Kadikis, and Kaspars Ozols. Perturbation-based methods for explaining deep neural networks: A survey. *Pattern Recognition Letters*, 150:228–234, 08 2021.

[30] Peng-Tao Jiang, Chang-Bin Zhang, Qibin Hou, Ming-Ming Cheng, and Yunchao Wei. Layercam: Exploring hierarchical class activation maps for localization. *IEEE Transactions on Image Processing*, 2021.

[31] Daniel S Kermany, Michael Goldbaum, Wenjia Cai, Carolina CS Valentim, Huiying Liang, Sally L Baxter, Alex McKeown, Ge Yang, Xiaokang Wu, Fangbing Yan, et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. *cell*, 172(5):1122–1131, 2018.

[32] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International conference on machine learning*, pages 2668–2677. PMLR, 2018.

[33] Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *International conference on machine learning*, pages 5338–5348. PMLR, 2020.

[34] Narine Kokhlikyan, Vivek Miglani, Miguel Martin, Edward Wang, Bilal Alsallakh, Jonathan Reynolds, Alexander Melnikov, Natalia Kliushkina, Carlos Araya, Siqi Yan, et al. Captum: A unified and generic model interpretability library for pytorch. *arXiv preprint arXiv:2009.07896*, 2020.

[35] Zewen Li, Fan Liu, Wenjie Yang, Shouheng Peng, and Jun Zhou. A survey of convolutional neural networks: analysis, applications, and prospects. *IEEE transactions on neural networks and learning systems*, 33(12):6999–7019, 2021.

[36] TorchVision maintainers and contributors. Torchvision: Pytorch's computer vision library. `https://github.com/pytorch/vision`, 2016.

[37] Joshua T Mattila, Michael J Fine, Andrew H Limper, Patrick R Murray, Bill B Chen, and Philana Ling Lin. Pneumonia. treatment and diagnosis. *Annals of the American Thoracic Society*, 11(Supplement 4):S189–S192, 2014.

[38] Sarah Mueller, Holger Heidrich, Lisa M. Koch, and Philipp Berens. fundus circle cropping.

[39] Ian E Nielsen, Ravi P Ramachandran, Nidhal Bouaynaya, Hassan M Fathallah-Shaykh, and Ghulam Rasool. Evalattai: a holistic approach to evaluating attribution maps in robust and non-robust models. *IEEE Access*, 11:82556–82569, 2023.

[40] Maxime Oquab, Léon Bottou, Ivan Laptev, and Josef Sivic. Is object localization for free?-weakly-supervised learning with convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 685–694, 2015.

[41] 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.

[42] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

[43] Lassi Raatikainen and Esa Rahtu. The weighting game: Evaluating quality of explainability methods. *arXiv preprint arXiv:2208.06175*, 2022.

[44] Harish Guruprasad Ramaswamy et al. Ablation-cam: Visual explanations for deep convolutional network via gradient-free localization. In *proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 983–991, 2020.

[45] 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.

[46] Adriel Saporta, Xiaotong Gui, Ashwin Agrawal, Anuj Pareek, Steven QH Truong, Chanh DT Nguyen, Van-Doan Ngo, Jayne Seekins, Francis G Blankenberg, Andrew Y Ng, et al. Benchmarking saliency methods for chest x-ray interpretation. *Nature Machine Intelligence*, 4(10):867–878, 2022.

[47] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[48] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Workshop Track Proceedings*, 2014.

[49] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[50] Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825*, 2017.

[51] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. Striving for simplicity: The all convolutional net. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Workshop Track Proceedings*, 2014.

[52] Susu Sun, Stefano Woerner, Andreas Maier, Lisa M Koch, and Christian F Baumgartner. Inherently interpretable multi-label classification using class-specific counterfactuals. In *Medical Imaging with Deep Learning*, 2023.

[53] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR, 2017.

[54] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

[55] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019.

[56] Haofan Wang, Zifan Wang, Mengnan Du, Fan Yang, Zijian Zhang, Sirui Ding, Piotr Mardziel, and Xia Hu. Score-cam: Score-weighted visual explanations for convolutional neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 24–25, 2020.

[57] Wenli Yang, Yuchen Wei, Hanyu Wei, Yanyu Chen, Guan Huang, Xiang Li, Renjie Li, Naimeng Yao, Xinyi Wang, Xiaotong Gu, Muhammad Amin, and Byeong Kang. Survey on explainable ai: From approaches, limitations and applications aspects. *Human-Centric Intelligent Systems*, 3:161–188, 2023.

[58] Hanwei Zhang, Felipe Torres Figueroa, and Holger Hermanns. Saliency maps give a false sense of explainability to image classifiers: An empirical evaluation across methods and metrics. In *The 16th Asian Conference on Machine Learning (Conference Track)*, 2024.

[59] Hanwei Zhang, Felipe Torres, Ronan Sicre, Yannis Avrithis, and Stephane Ayache. Opti-cam: Optimizing saliency maps for interpretability. *Computer Vision and Image Understanding*, 248:104101, 2024.

[60] Jianming Zhang, Sarah Adel Bargal, Zhe Lin, Jonathan Brandt, Xiaohui Shen, and Stan Sclaroff. Top-down neural attention by excitation backprop. *International Journal of Computer Vision*, 126(10):1084–1102, 2018.

[61] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016.

[62] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 67(2):301–320, 2005.

A Implementation Details

A.1 Datasets

We evaluate our approach on three publicly available medical imaging datasets spanning three different modalities: the Kaggle Diabetic Retinopathy (DR) [20], the Retinal OCT dataset [31], and the RSNA Chest X-ray (CXR) dataset [1].

- **Kaggle DR Dataset.** This dataset comprises $88,702$ high-resolution retinal fundus images labeled for DR severity on a 5-point scale from 0 (No DR) to 4 (Proliferative DR). After applying an automated quality filtering pipeline using an ensemble of EfficientNet models [55] trained on the ISBI2020[2] challenge dataset, we retained $45,923$ images from $28,984$ subjects. The resulting class distribution was 73% (class 0), 15%, 8%, 3%, and 1%. For binary classification (early DR detection), we grouped class $\{0\}$ vs. $\{1,2,3,4\}$, yielding an imbalance of 73% vs. 27%. Additionally, lesion annotations for 65 images were obtained from [17] for evaluating the model's explanations at localizing DR-related lesions.
- **Retinal OCT Dataset.** This dataset consists of $108,315$ B-scans categorized into four classes: Drusen, Diabetic macular edema (DME), Choroidal neovascularization (CNV), and Normal. A separate test set of $1,000$ B-scans is provided. Following [16], we excluded low-resolution scans (width ≤ 496). As preliminary experiments showed that using the full dataset did not significantly improve performance, we subsampled the training set (by randomly removing half of the healthy images [16]) to $34,962$ scans ($8,616$ Drusen, $26,346$ Normal) for binary classification (Drusen vs. Normal), preserving the original class imbalance (73% vs 27%). Additionally, we used 40 drusen-annotated B-scans from [16] to evaluate the model's explanations at localizing drusen lesions. For the multi-class classification task, the training was randomly reduced to $17,200$ images while maintaining the original class distribution: 45% Normal, 34% CNV, 10% DME, and 9% Drusen.
- **RSNA Chest X-ray Dataset.** This dataset includes $30,227$ frontal-view chest radiographs labeled as "Normal", "No Opacity/Not Normal", and "Opacity" (indicative of pneumonia). Pneumonia cases come with bounding box annotations, which facilitate the evaluation of the model's explanations. For our binary classification task, we selected images labeled as either "Normal" or "Opacity", resulting in $14,863$ images with a 60% vs. 40% class distribution.

Each dataset was split into training (75%), validation (10%), and test (15%) sets, except for the Retinal OCT dataset, which followed an 80%-20% training-validation split, due to its predefined test set (250 images per class). All training splits used in our experiments are provided in CSV format and publicly available via the project's GitHub[3] repository.

A.2 Baseline models

The effectiveness of our method was evaluated using two widely adopted black-box CNN architectures: ResNet-50 [25] and VGG-16 [49]. These models were chosen due to their distinct architectures, such as depth, theoretical receptive field size, and classification head design, which allow for a broad assessment of our method's generalizability. In both models, the standard classification head was replaced with our proposed convolutional evidence map layer to enable inherent interpretability. For ResNet50, we removed the global average pooling layer and final fully connected layer, substituting them with a class evidence layer consisting of C convolutional filters (1×1, stride 1), where C is the number of output classes. This layer directly produces class-specific evidence maps (Sec. 3.2).

For VGG-16, which uses a series of fully connected layers (FCLs) in its classifier head, each FCL was replaced by an equivalent 1×1 convolutional layer. Specifically, an FCL of size $b_1 \times b_2$ was transformed into a convolutional layer of size $b_1 \times b_2 \times 1\ times1$, preserving the original parameter count and model capacity. These architectural adjustments maintain model complexity and capacity while introducing interpretability directly into the classification mechanism.

[2] `https://isbi.deepdr.org/challenge2.html`

[3] Code and CSV files are available at `https://anonymous.4open.science/r/SoftCAM-E1A3/`

A.3 Data preprocessing

Fundus images were preprocessed by cropping them to a square shape using a circle-fitting method as described in [38]. All datasets were then resized to 512×512 pixels, except for the retinal OCT dataset, which was resized to 496×496 to better match its original lower resolution. Image intensities were normalized using the mean and standard deviation computed from the respective training sets.

During training, consistent data augmentation strategies were applied across all datasets. These included flipping, rotation, random cropping, and translation, each applied with a fixed probability. For the Kaggle dataset, which contains color fundus images, additional color augmentations were introduced to improve generalization.

A.4 Training setup

All models were obtained from Torchvision and initialized with pretrained ImageNet weights. They were subsequently fine-tuned on each dataset using a consistent training setup. Following [16, 18], we employed the cross-entropy loss function and optimized model parameters using stochastic gradient descent (SGD) with Nesterov momentum (momentum factor of 0.9). The initial learning rate was set to 1.10-3, and a clipped cosine annealing learning rate scheduling was applied with the minimum learning rate set to 1.10-4. Weight decay was set to 5.10-4. Training was conducted for 70 epochs with a mini-batch size of 16 on an NVIDIA A40 GPU using PyTorch [42].

A.5 Baseline CAM-based methods

Gradient-based methods primarily differ in how gradients are aggregated to compute importance weights, while gradient-free methods mainly vary in how the weights are computed.

ScoreCAM [56]. A gradient-free method that eliminates the need for gradient information by assessing the importance of each activation map based on its forward-pass contribution to the target class score, and produces the final output via a weighted sum of these maps.

LayerCAM [30]. A gradient-based method that generates class activation maps by leveraging the element-wise product of ReLU-activated gradients and feature maps at any convolutional layer, enabling fine-grained, spatially precise visual explanations without requiring global average pooling.

GradCAM [47]. A gradient-based approach that uses the gradients of the target class flowing into the final convolutional layer to produce a coarse localization map, highlighting important regions in the image by upsampling the resulting map.

Guided backpropagation (Guided BP) [51]. A gradient-based approach that modifies the standard backpropagation process to propagate only positive gradients through positive activations, producing fine-grained visualizations that highlight features strongly activating specific neurons in relation to the target output.

Integrated Gradient (Itgt Grad.) [53]. A gradient-based method that attributes model predictions to input features by computing the path integral of gradients along a straight-line path from a baseline to the actual input, yielding fine-grained explanations.

B Explainability metrics

B.1 Activation consistency

The activation consistency [18] quantifies how well local explanations (e.g., individual activations within explanation maps) globally reflect the disease and healthy samples across a dataset. Specifically, it measures whether the activation patterns in the explanation maps consistently reflect the underlying disease or healthy class labels.

Following [18], we evaluated activation consistency by computing the proportion of positive activations (indicative of disease evidence) in saliency maps of disease samples, and negative activations

(indicative of the absence of disease) in those of healthy samples. These proportions were calculated over the test set to assess whether the heatmaps consistently highlight pathological features in diseased cases and suppress activations in healthy ones. This metric thus captures the alignment between the semantic meaning of activations and the ground truth labels, offering a dataset-level evaluation of the coherence of local explanations with the global classification objective.

B.2 Activation precision and activation sensitivity

Let $\mathcal{X} = \{\mathbf{X}\}_{\mathbf{i=1}}^{\mathbf{n}}$ denote a set of input images, $\mathcal{M} = \{\mathbf{M}\}_{\mathbf{i=1}}^{\mathbf{n}}$ the corresponding binary segmentation masks, and $\mathcal{S} = \{\mathbf{S}\}_{\mathbf{i=1}}^{\mathbf{n}}$ the associated explanation or saliency maps generated by any method. *Activation precision* measures the proportion of the saliency map's positive mass that lies within the annotated region (the segmentation mask) [9, 43]. To compute it, saliency maps are first preprocessed by setting negative values to zero while retaining all positive values. This highlights how much of the explanation signal aligns with human-annotated ground truth, effectively quantifying the precision of an explainability method. The activation precision is defined as:

$$AP(\mathcal{M}, \mathcal{S}) = \frac{\sum_{M,S} \sum_{i,j} M_{i,j}.S_{i,j}}{\sum_{i,j} S_{i,j}}. \tag{6}$$

However, activation precision does not penalize false-negative (i.e. missed relevant regions). To address this, we introduce *activation sensitivity*, which captures the completeness of the explanation by evaluating how much of the annotated region is covered by the saliency map. The activation precision is defined as:

$$AS(\mathcal{M}, \mathcal{S}) = \frac{\sum_{M,S} \sum_{i,j} M_{i,j}.S_{i,j}}{\sum_{i,j} M_{i,j}}. \tag{7}$$

Unlike activation precision, activation sensitivity penalizes low saliency values within the mask. For example, if $M_{i,j} = 1$ but $0 < S_{i,j} < 1$, the low activation will contribute little to the numerator, reflecting reduced confidence in that region. This makes activation sensitivity especially relevant in clinical tasks where completeness is critical, such as identifying multi-focal infectious diseases like pneumonia [37].

B.3 Top-k localization precison

Top-k localization precision [18] measures the ability of an explanation map to correctly highlight salient regions that overlap with ground-truth annotations. Specifically, it quantifies the proportion of the top-k positively activated regions within an explanation that match with annotated areas. In our implementation, each explanation map is first upsampled to the input resolution and then split into non-overlapping patches of size 33×33. These patches are ranked based on their average activation, and the top-k (k=30) most salient patches are selected. The precision is then computed as the fraction of these patches that overlap with the annotated ground-truth regions.

This metric can be viewed as a generalization of the pointing game metric [60], where only the single most activated region (top-1) is considered, to multiple regions, making it more suitable for medical imaging tasks. In such contexts, disease-relevant features (e.g., retinal lesions or pathological markers) are often spatially distributed across the image, rather than confined to a single localized area.

B.4 Failthfulness

Faithfulness, also referred to as sensitivity or fidelity [39], is a widely used metric to evaluate how accurately an explanation reflects the model's true decision-making process. It assesses whether the importance scores (attributions) assigned to input features correspond to the actual impact of those features on the model's prediction.

In our implementation, we focus on correctly classified samples from the test set. For each, the corresponding explanation map is upsampled to the input resolution and split into non-overlapping patches of size 33×33. These patches are ranked based on their mean activation values, and the top-k (k=30) most salient patches are iteratively occluded. After each occlusion step, we recorded the relative drop in the model's confidence score for the predicted class. This process yields a

deletion curve, from which we compute the Area Under the Deletion Curve (AUDC). A lower AUDC indicates a more faithful explanation, as it reflects a greater decline in model confidence when the most important regions (as indicated by the explanation map) are removed, suggesting that those regions were indeed critical to the model's prediction.

C Effect of Lasso regularization on model performance for the binary tasks

The Lasso regularization coefficient λ_1 in Eq. 5 controls the sparsity of the class evidence map, encouraging the model to localize disease regions with high precision. For each task, λ_1 was selected based on a trade-off between accuracy and AUC on the corresponding validation set, choosing the highest values for which classification performance did not degrade significantly.

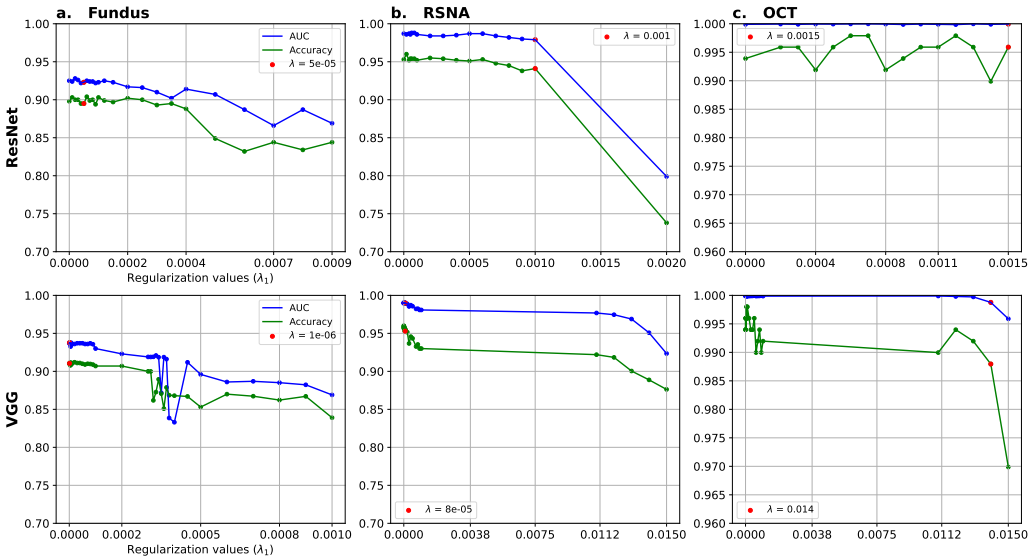

Figure 6: **Model selection on validation sets under varying Lasso regularization strengths.** The regularization coefficient λ influences model performance, with notable effects on some datasets but minimal impact on the OCT dataset. The red markers indicate the selected λ values, chosen to balance sparsity and classification performance.

D Additionnal Results

D.1 SoftCAM provides inherently interpretable visual explanations

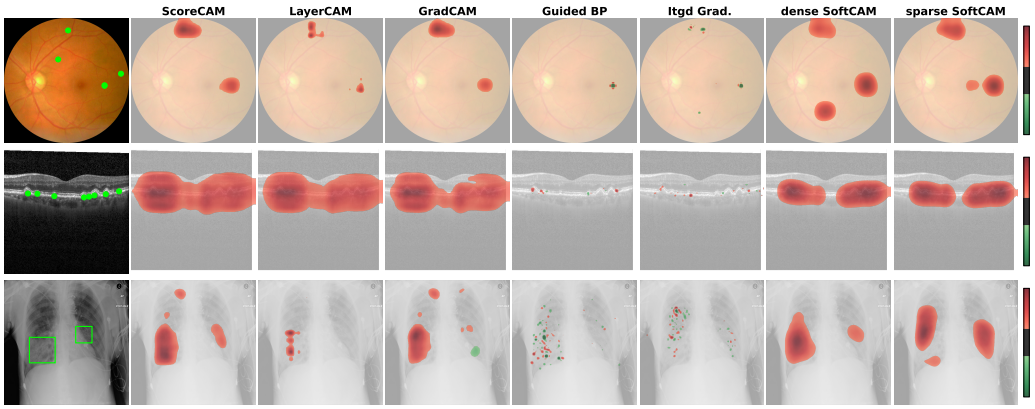

Figure 7: **Example explanations generated by different methods from VGG-16**. The first column shows disease images with reference annotations, indicated by green markers or bounding boxes. Each row, from top to bottom, corresponds to fundus, OCT, and Chest X-ray images, respectively. The next five columns present saliency maps generated by post-hoc explanation methods, gradient-free (ScoreCAM, LayerCAM) and gradient-based (GradCAM, Guided BP, Itgd Grad). The final two columns showcase our proposed inherently interpretable dense and sparse SoftCAM explanations.

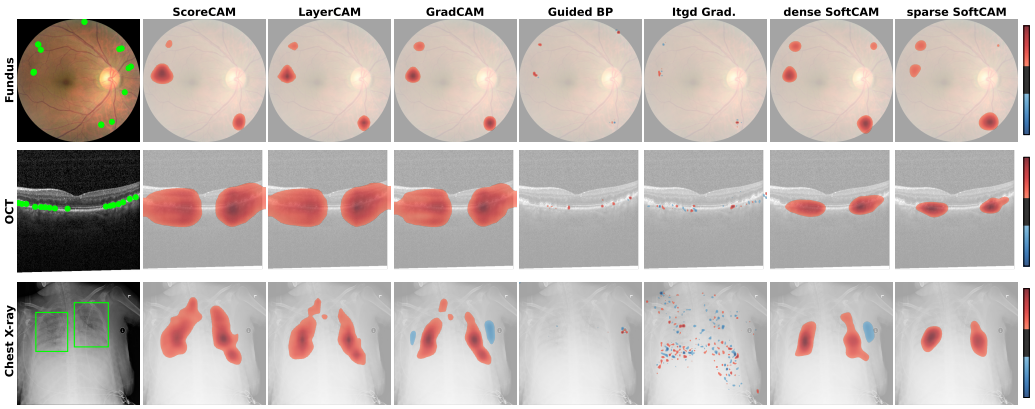

Figure 8: **Additional example explanations of disease images from the ResNet model**.

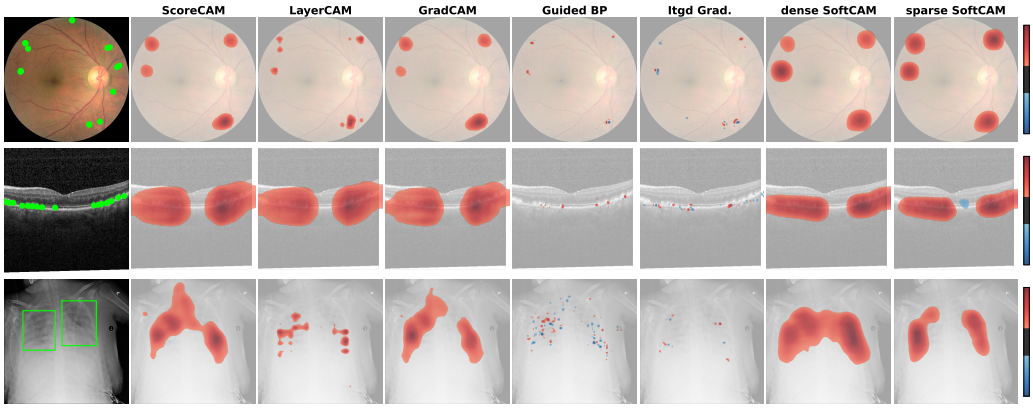

Figure 9: **Additional example explanations of disease images from the VGG model**.

D.2 Activation consistency

We quantify activation consistency only for the SoftCAM variants, as post-hoc methods are not inherently explainable, meaning their explanations do not directly influence the model's decision-making process.

The results align well with qualitative visualizations. On the Fundus dataset, the sparse SoftCAM model exhibits a higher proportion of positive activations with the ResNet backbone, attributed to reduced false positives from the dense model, and fewer negative activations, reflecting the suppression of low-importance activations to zero. On the VGG backbone, regularization primarily reduces false-positive activations from the dense model but leads to a slight increase in activations on healthy samples. Similar result can be observed on the RSNA dataset.

On the OCT dataset, the dense SoftCAM with the ResNet backbone generally produces coarse-grained evidence around lesion areas. In contrast, the sparse variant refines these explanations, resulting in lower positive and negative activations across both disease and healthy samples, suggesting more selective and focused localization. However, with the VGG backbone, a higher proportion of negative activations is observed, reflecting the impact of the regularization strength, highlighting the importance of appropriately tuning this parameter for different architectures.

Table 2: Activation consistency on the ResNet model. $\mathbf{r}^{+}_{\mathbf{LG}}$ denotes the proportion of positive or disease activations from disease images, while $\mathbf{r}^{-}_{\mathbf{LG}}$ refers to the proportion of negative or healthy activations from healthy images.

	Fundus		OCT		RSNA	
	$\mathbf{r}^{+}_{\mathbf{LG}} \uparrow$	$\mathbf{r}^{-}_{\mathbf{LG}} \uparrow$	$\mathbf{r}^{+}_{\mathbf{LG}} \uparrow$	$\mathbf{r}^{-}_{\mathbf{LG}} \uparrow$	$\mathbf{r}^{+}_{\mathbf{LG}} \uparrow$	$\mathbf{r}^{-}_{\mathbf{LG}} \uparrow$
dense SoftCAM	0.28 ± 0.1	0.86 ± 0.1	0.30 ± 0.1	0.85 ± 0.1	0.75 ± 0.1	0.47 ± 0.1
sparse SoftCAM	0.55 ± 0.2	0.76 ± 0.2	0.23 ± 0.1	0.83 ± 0.1	0.79 ± 0.1	0.45 ± 0.1

Table 3: Activation consistency on the VGG model. $\mathbf{r}^{+}_{\mathbf{LG}}$ denotes the proportion of positive or disease activations from disease images, while $\mathbf{r}^{-}_{\mathbf{LG}}$ refers to the proportion of negative or healthy activations from healthy images.

	Fundus		OCT		RSNA	
	$\mathbf{r}^{+}_{\mathbf{LG}} \uparrow$	$\mathbf{r}^{-}_{\mathbf{LG}} \uparrow$	$\mathbf{r}^{+}_{\mathbf{LG}} \uparrow$	$\mathbf{r}^{-}_{\mathbf{LG}} \uparrow$	$\mathbf{r}^{+}_{\mathbf{LG}} \uparrow$	$\mathbf{r}^{-}_{\mathbf{LG}} \uparrow$
dense SoftCAM	0.32 ± 0.2	0.93 ± 0.1	0.75 ± 0.11	0.51 ± 0.1	0.75 ± 0.1	0.51 ± 0.1
sparse SoftCAM	0.28 ± 0.2	0.94 ± 0.1	0.35 ± 0.14	0.95 ± 0.1	0.35 ± 0.1	0.95 ± 0.1

Overall, the effect of regularization on the explanations varies depending on the backbone architecture. Nevertheless, the activation consistency metric aligns well with the qualitative explanations, generally capturing the impact of regularization across the dataset for a given architecture.

D.3 Precision and sensitivity analysis

We quantitatively evaluate the explanations generated by various methods using the ResNet and VGG backbones on the RSNA dataset. With the ResNet model, the dense SoftCAM achieves the highest localization precision, whereas the sparse SoftCAM yields the best results in terms of sensitivity. This discrepancy underscores the importance of developing evaluation metrics that balance human-aligned localization quality with model fidelity, capturing both interpretability and decision relevance.

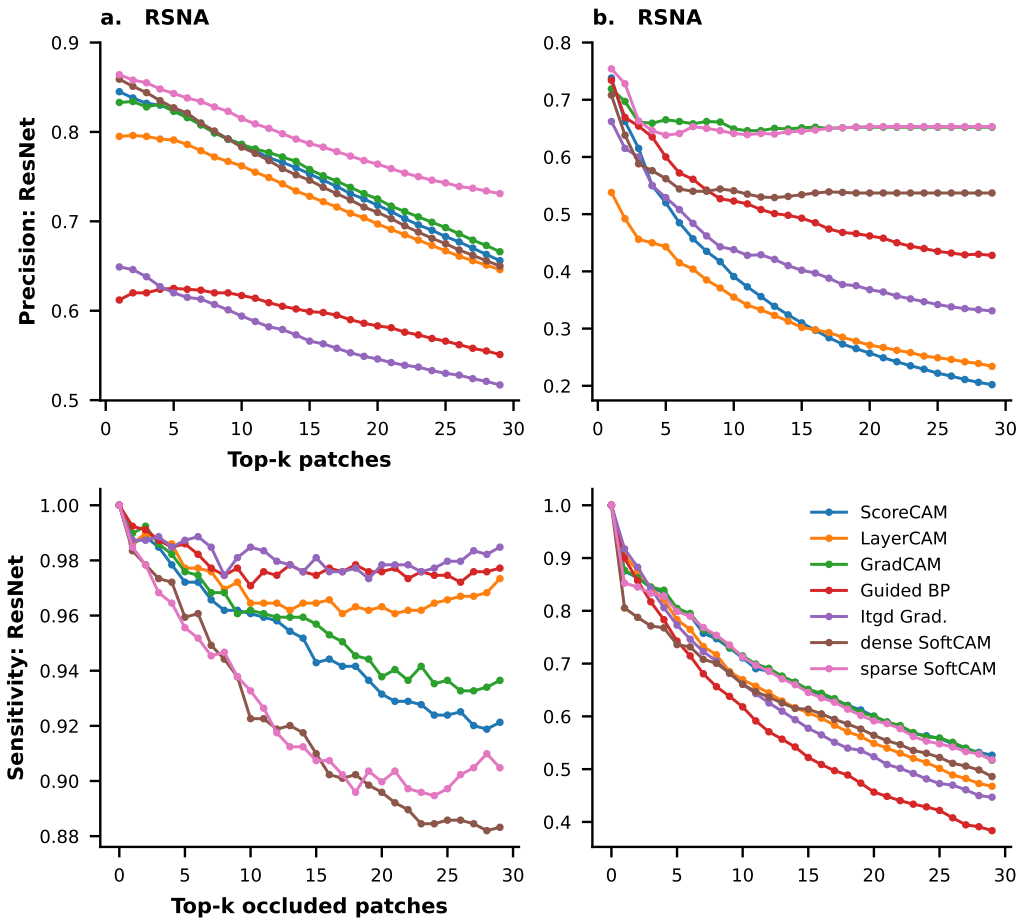

Figure 10: **Precision vs. sensitivity analysis on the RSNA dataset**. Quantitative evaluation of explanations generated by different methods from the ResNet and VGG models on the RSNA dataset.

D.4 SoftCAM provides localized and faithful explanations

Table 4: Top-k localization precision and sensitivity. Sensitivity is quantified as the Area Under the Deleted Curve (AUDC), where lower values indicate greater faithfulness—that is, a larger drop in the model's confidence when the most relevant patches are removed. For precision, higher values indicate better alignment between saliency maps and ground truth annotations. We refer to AUDC as "Del" and Top-K as "Top" with $K = 30$.

	ResNet (Topk ↑, AUDC ↓)						VGG (Topk ↑, AUDC ↓)					
	Fundus		OCT		RSNA		Fundus		OCT		RSNA	
	Top	Del	Top	Del	Top	Del	Top	Del	Top	Del	Top	Del
ScoreCAM	0.16	0.67	0.07	0.73	0.66	0.97	0.20	0.67	0.08	0.55	0.62	0.88
LayerCAM	0.22	0.65	0.08	0.74	0.65	0.97	0.23	0.64	0.08	0.56	**0.65**	0.84
GradCAM	0.37	0.64	0.14	0.73	0.67	0.95	0.65	0.68	0.14	0.58	0.61	0.86
Guided BP	0.30	**0.57**	0.23	0.68	0.55	0.97	0.43	0.57	0.23	**0.40**	0.58	0.85
Itgd Grad.	0.28	0.63	0.20	0.70	0.52	0.98	0.33	**0.62**	0.2	0.51	0.55	0.88
dense SoftCAM	0.39	0.69	0.46	0.61	0.65	**0.92**	0.54	0.63	0.72	0.45	0.64	0.84
sparse SoftCAM	**0.52**	0.68	**0.86**	**0.31**	**0.73**	0.93	**0.65**	0.674	**0.82**	0.43	0.63	**0.82**

E Activation precision and sensitivity on the RSNA dataset

E.1 Lasso vs Ridge penalty

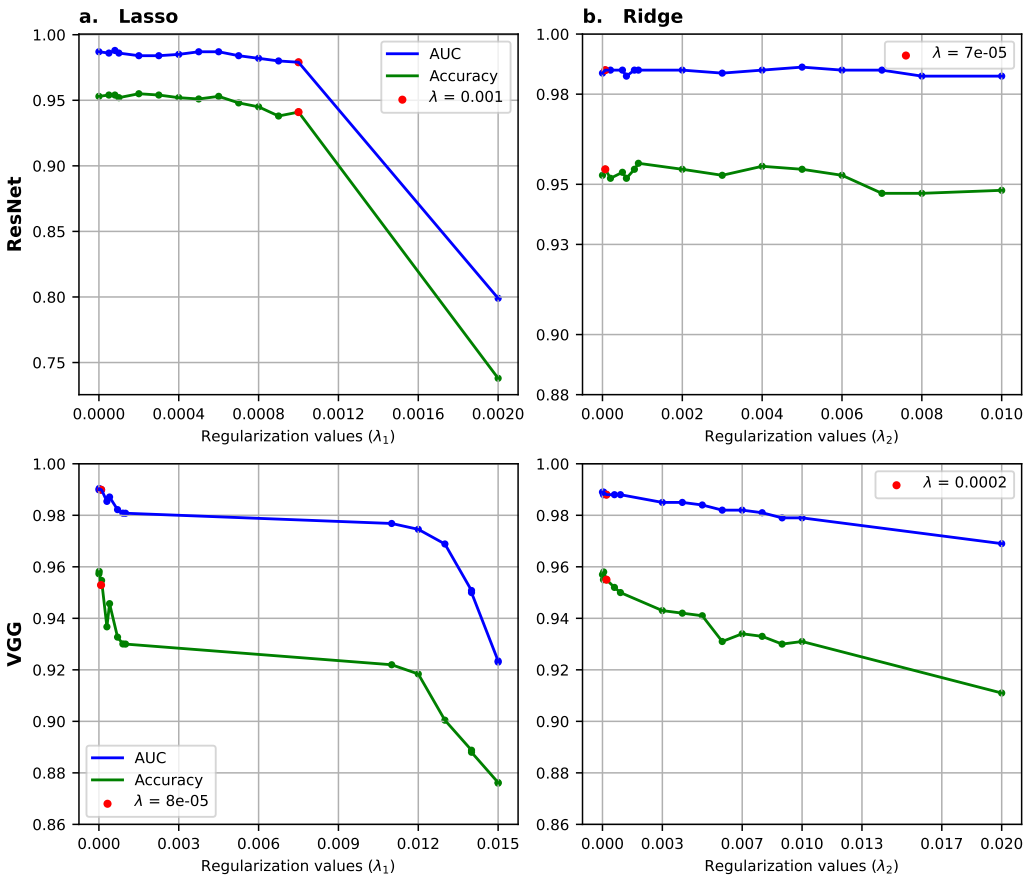

Figure 11: **Model selection on validation sets under varying Lasso and Ridge regularization strengths.** The regularization coefficients λ_1 and λ_2 influence model performance. The red markers indicate the selected regularization values, chosen to balance classification performance.

E.2 Activation precision vs. activation sensitivity

Table 5: Activation Precision (AP) vs. Activation Sensitivity (AS) for different SoftCAM variants and baseline post-hoc methods. The dense SoftCAM consistently lies between the lasso and ridge variants, highlighting the importance of balancing the two regularization terms to achieve an optimal trade-off between precision and completeness in the explanations.

	ResNet		VGG	
	AP ↑	AS ↑	AP ↑	AS ↑
ScoreCAM	0.470	0.318	0.403	0.303
LayerCAM	0.456	0.300	0.401	0.120
GradCAM	0.525	0.252	0.373	0.260
Guided BP	0.381	0.033	0.364	0.044
Itgd Grad.	0.286	0.040	0.322	0.039
dense SoftCAM	0.526	0.251	0.461	0.355
sparse SoftCAM	**0.654**	0.182	**0.519**	0.320
lasso SoftCAM	0.440	**0.316**	0.412	**0.396**

E.3 More examples: activation precision vs activation sensitivity

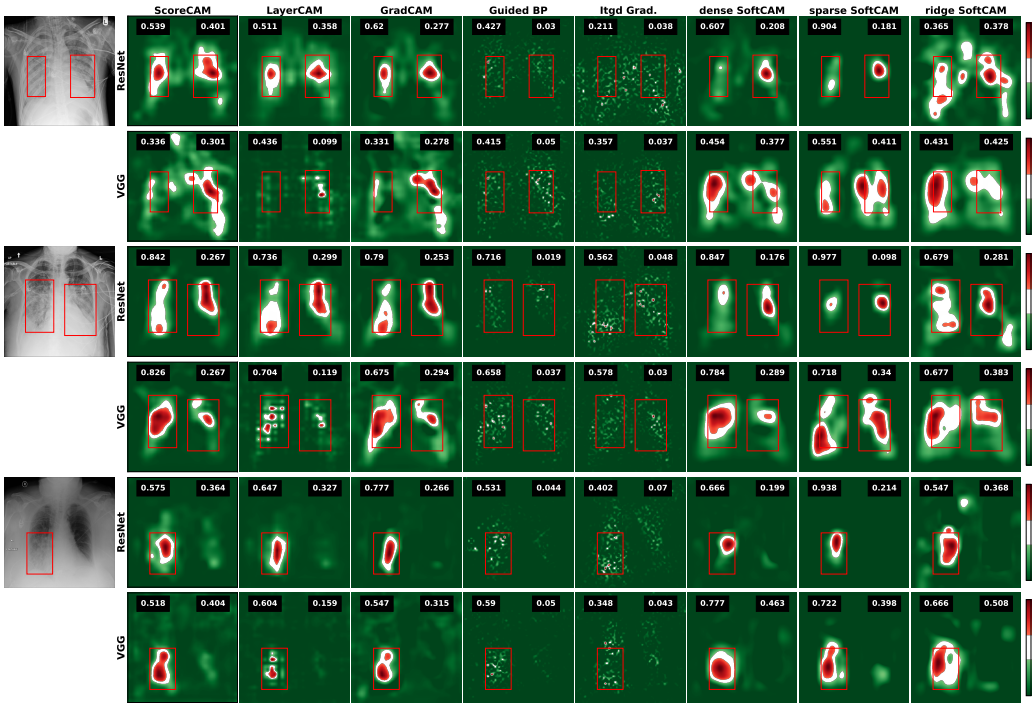

Figure 12: **Additional examples of localization evaluation on the RSNA dataset for pneumonia detection**. Each column shows explanation maps generated by different methods. Ground-truth bounding boxes are overlaid on each map, with the top-right value indicating the activation precision, while the top-left value indicates the activation sensitivity. The high precision from the lasso model and the more complete explanations from the ridge model emphasize the importance of balancing the two regularization terms to achieve an optimal trade-off.

F Multi-class analysis

F.1 Regularization

Given the small size of retinal lesions, we used Lasso regularization, selecting λ_1 values that balanced performance (Fig. 13)

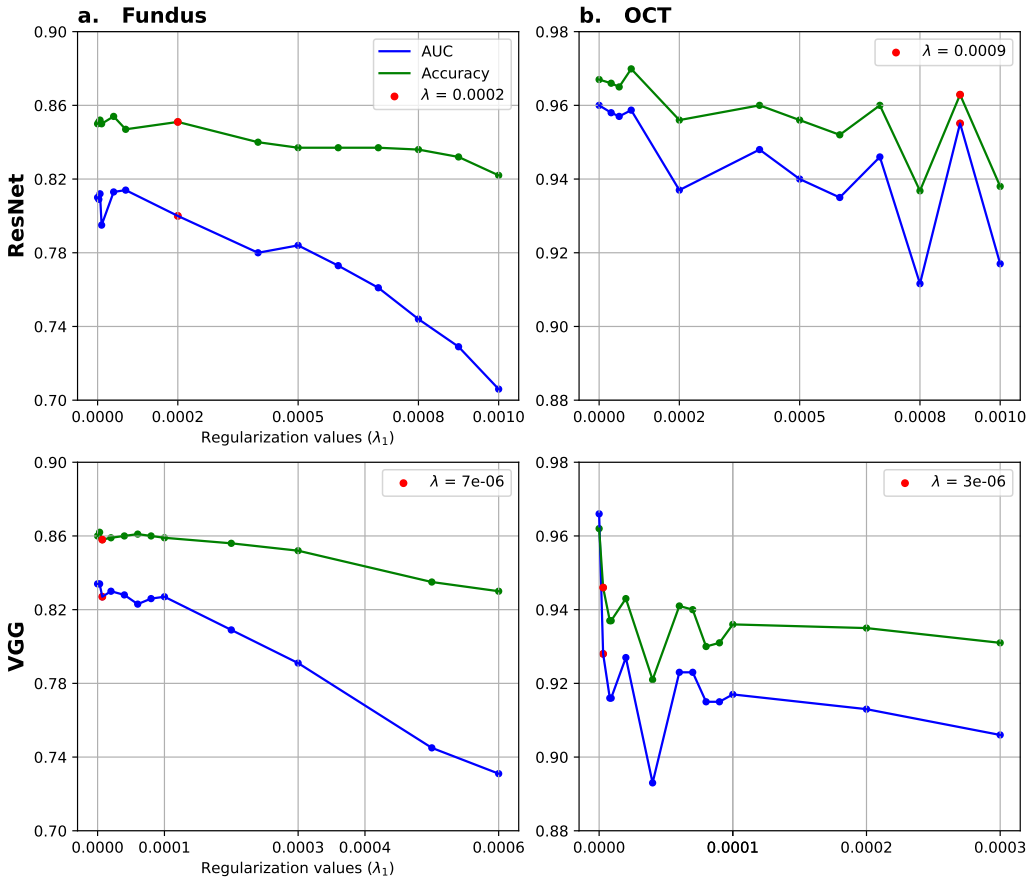

Figure 13: **Model selection on validation sets under varying Lasso regularization strengths.** The regularization coefficients λ_1 influence model performance. The red markers indicate the selected regularization values to balance classification performance.

F.2 Faithfulness

As no ground-truth lesion annotations were available for the multi-class tasks, we evaluated the faithfulness of the explanations by measuring their contribution to model predictions. For correctly classified test samples, we progressively removed top-k ($k = 30$) ranked patches (based on the explanation maps) and tracked the average drop in class confidence (Fig. 14).

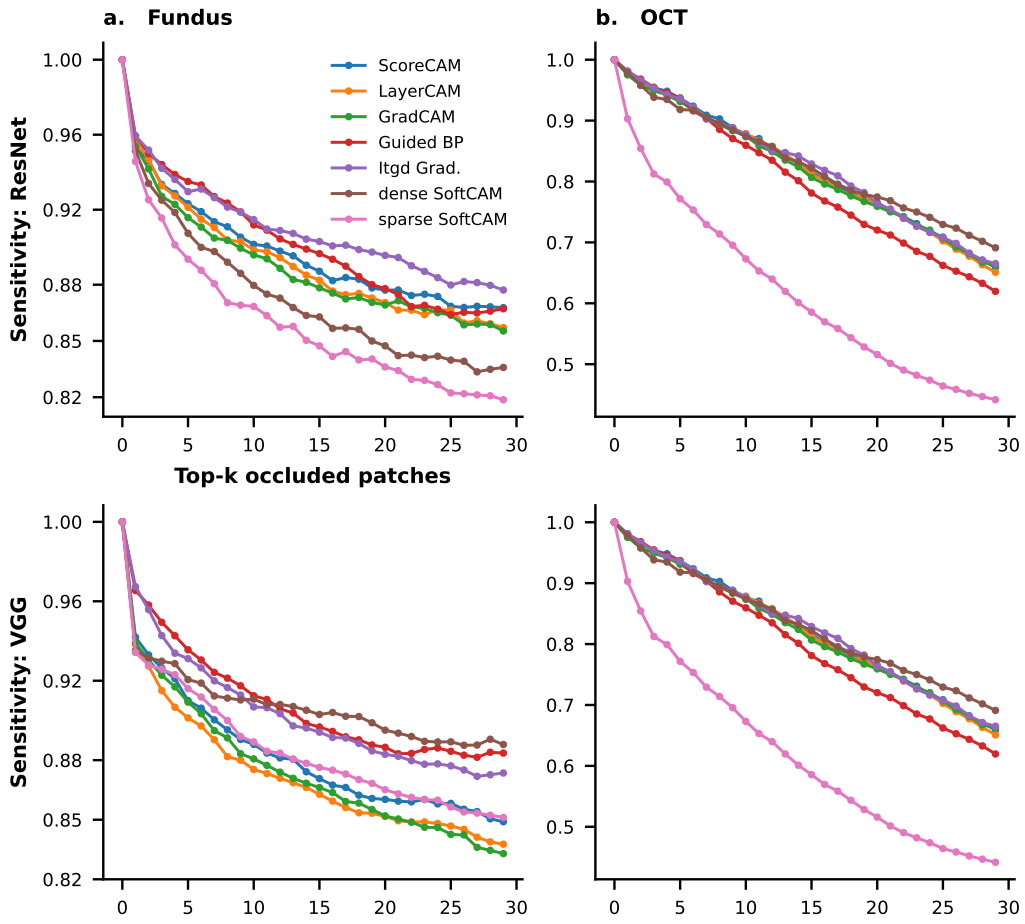

Figure 14: **Sensitivity analysis**.

F.3 Area Under the Deleted Curve

The area under the deletion curve (AUDC) was computed from the sensitivity analysis (Fig. 14). In both tasks, the dense and sparse SoftCAM achieved superior performance, with sparse SoftCAM yielding the lowest AUDC, indicating the highest faithfulness (Tab. 6).

Table 6: Area Under the Deleted Curve (AUDC ↓).

	ResNet		VGG	
	Fundus	OCT	Fundus	OCT
ScoreCAM	0.894	0.819	0.880	0.852
LayerCAM	0.889	0.817	**0.869**	0.850
GradCAM	0.887	0.815	0.872	0.847
Guided BP	0.899	0.793	0.905	0.823
Itgd Grad.	0.907	0.821	0.901	0.833
dense SoftCAM	0.870	0.825	0.905	0.826
sparse SoftCAM	**0.856**	**0.609**	0.882	**0.806**

F.4 Qualitative explanation on retinal fundus images

For the multi-class tasks on DR detection from fundus images, SoftCAM variants produced more focused and class-consistent explanations. In addition to the sparse and dense evidence maps, we also provide visualizations for post-hoc methods.

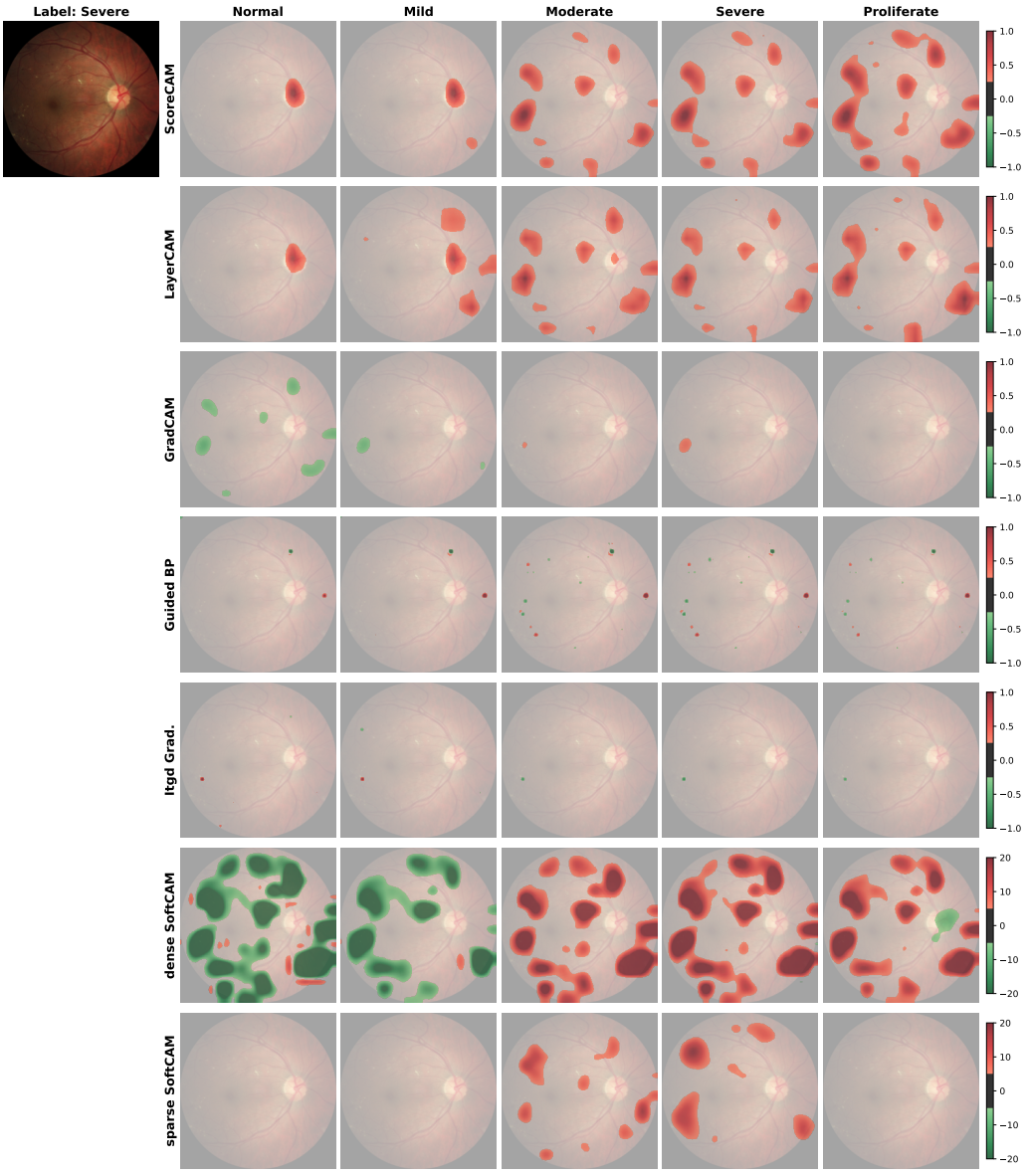

Figure 15: **Class-specific explanation with the ResNet backbone**. The application of our method to multi-class DR detection demonstrates the utility of class-specific explanations produced by the sparse SoftCAM, which more precisely highlight disease-relevant regions compared to the dense SoftCAM and the best-performing post-hoc method, GradCAM. In the example shown, the image is labeled as severe DR, and the highlighted regions correspond to suspicious areas, reflecting relevant DR lesions.

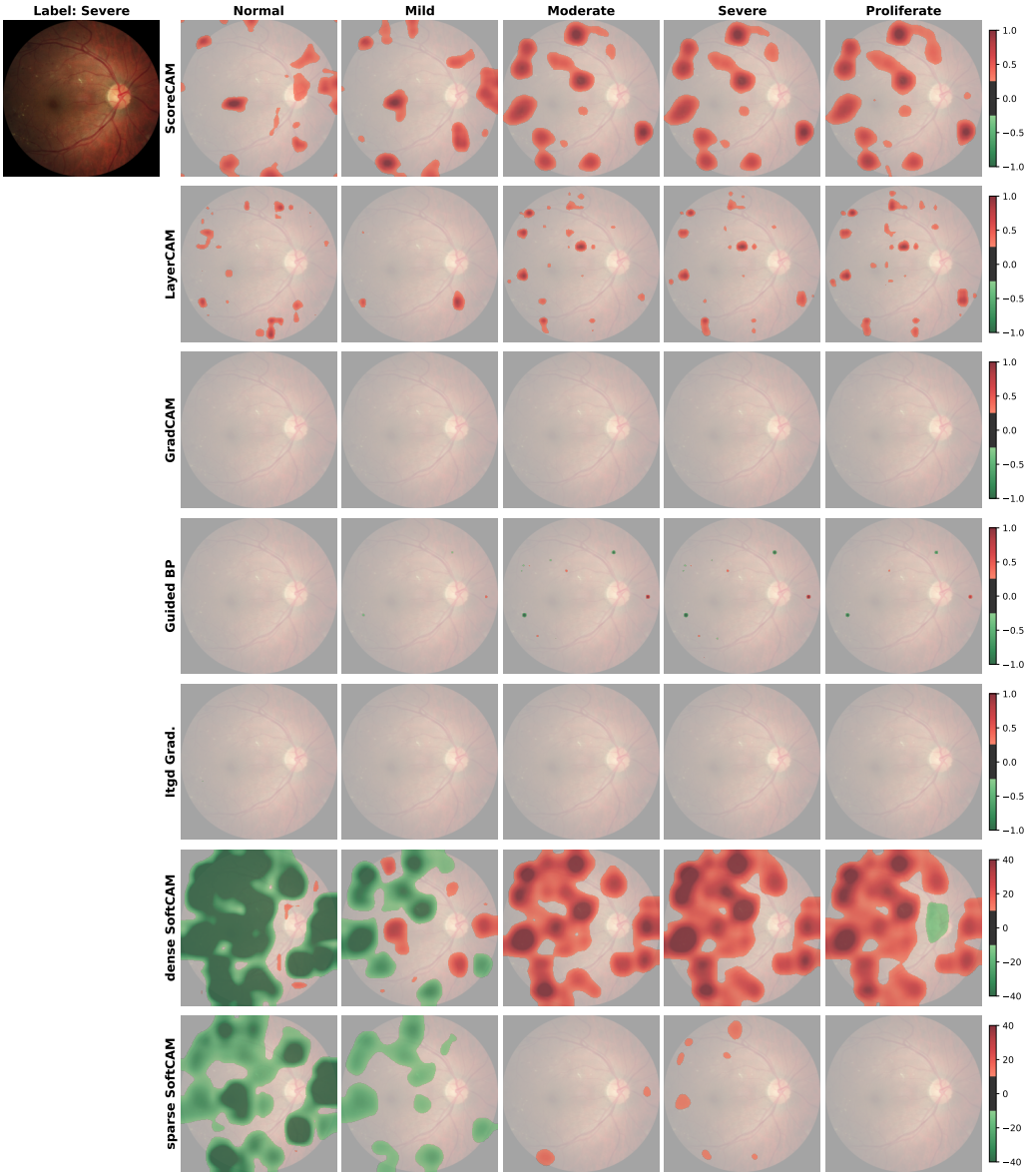

Figure 16: **Class-specific explanation with the VGG backbone**. The application of our method to multi-class DR detection demonstrates the utility of class-specific explanations produced by the sparse SoftCAM, which more precisely highlight disease-relevant regions compared to the dense SoftCAM and the best-performing post-hoc method, ScoreCAM. In the example shown, the image is labeled as severe DR, and the highlighted regions correspond to suspicious areas, reflecting relevant DR lesions.

F.5 Qualitative explanation on retinal OCT images

For the multi-class tasks on retinal disease classification from OCT images, SoftCAM variants produced more focused and class-consistent explanations. In addition to the sparse and dense evidence maps, we also provide visualizations for GradCAM and Guided BP, as these were the best-performing post-hoc methods for the ResNet and VGG backbones, in terms of the Area Under the Deletion Curve (Tab. 6).

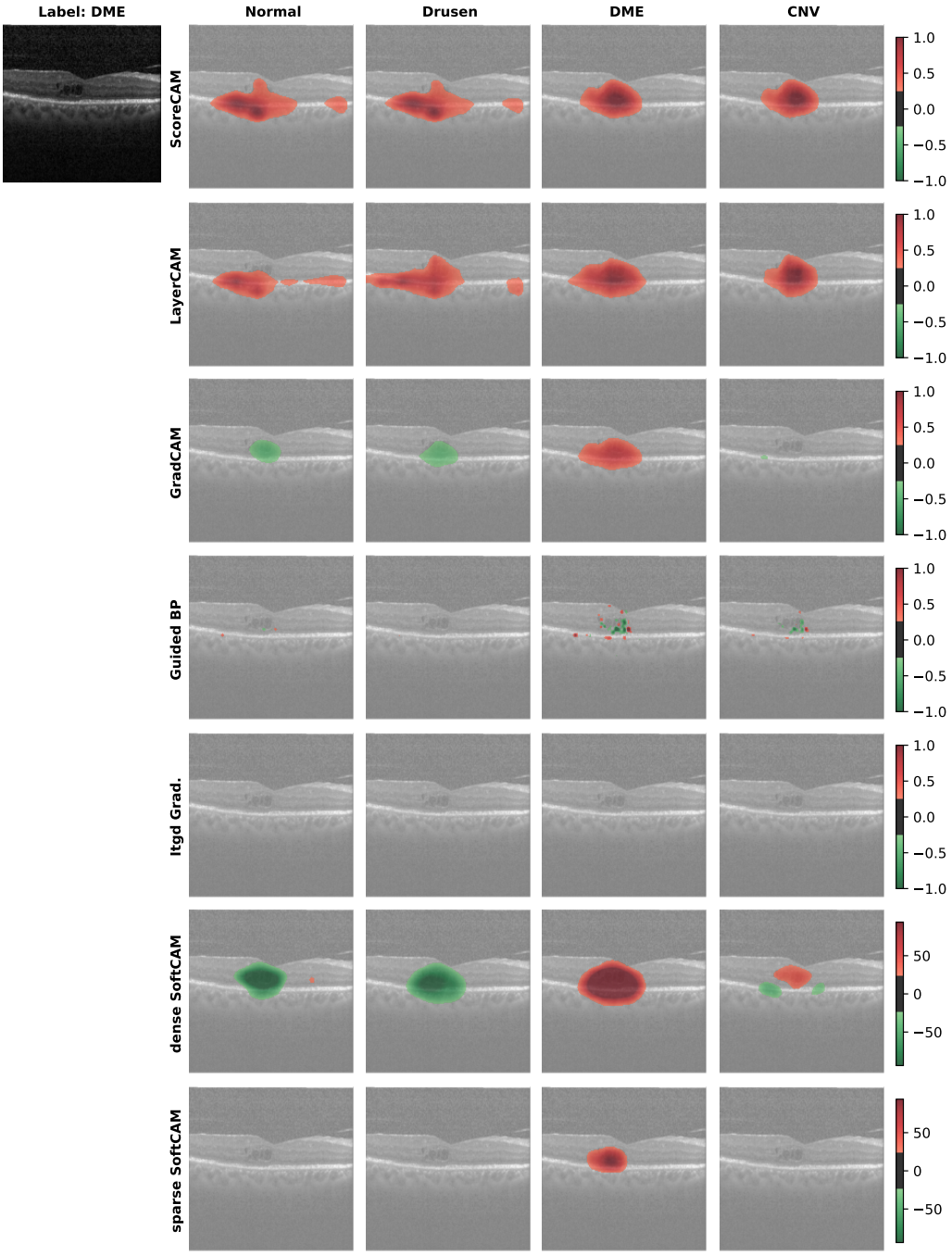

Figure 17: **Class-specific explanation with the ResNet backbone**. SoftCAM applied to multi-class retinal disease classification demonstrates the utility of class-specific explanations, with the sparse SoftCAM variant more precisely highlighting disease-relevant regions compared to both the dense SoftCAM and the best-performing post-hoc methods, GradCAM and Guided Backpropagation. In the example shown, the image is labeled as Diabetic Macular Edema (DME), and the highlighted regions produced by sparse SoftCAM highlight suspicious areas, reflecting relevant retinal lesions.

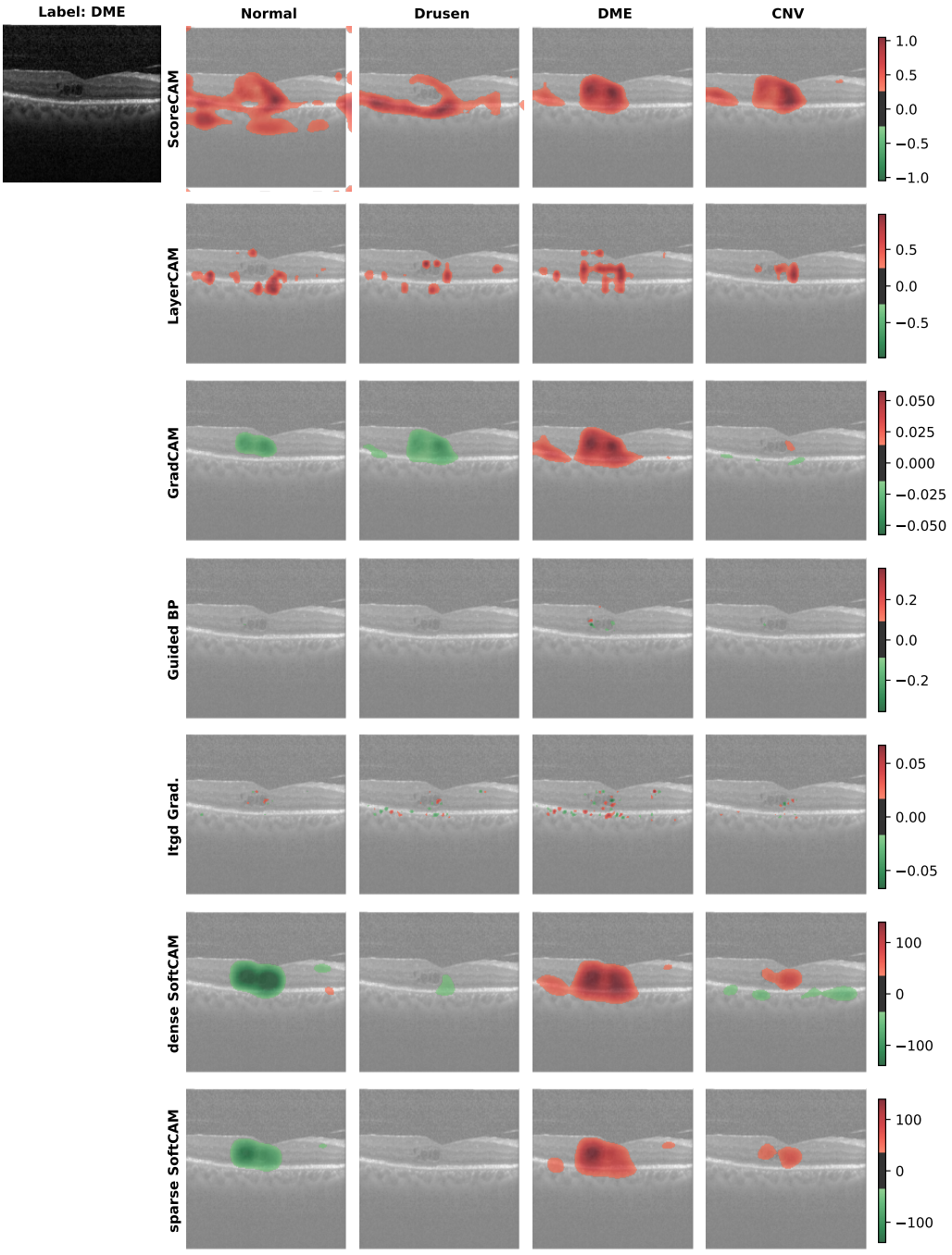

Figure 18: **Class-specific explanation with the VGG backbone**. SoftCAM applied to multi-class retinal disease classification demonstrates the utility of class-specific explanations, with the sparse SoftCAM variant more precisely highlighting disease-relevant regions compared to both the dense SoftCAM and the best-performing post-hoc methods, GradCAM and Guided Backpropagation. In the example shown, the image is labeled as Diabetic Macular Edema (DME), and the highlighted regions produced by sparse SoftCAM highlight suspicious areas, reflecting relevant retinal lesions.

A Hybrid Fully Convolutional CNN-Transformer Model for Inherently Interpretable Disease Detection from Retinal Fundus Images

Kerol Djoumessi[1(✉)[0009−0004−1548−9758]], Samuel Ofosu Mensah[1[0000−0002−9290−1206]], and Philipp Berens[1,2(✉)[0000−0002−0199−4727]]

[1] Hertie Institute for AI in Brain Health, University of Tübingen, Germany
{kerol.djoumessi-donteu, philipp.berens}@uni-tuebingen.de
[2] Tübingen AI Center, University of Tübingen, Germany

Abstract. In many medical imaging tasks, convolutional neural networks (CNNs) efficiently extract local features hierarchically. More recently, vision transformers (ViTs) have gained popularity, using self-attention mechanisms to capture global dependencies, but lacking the inherent spatial localization of convolutions. Therefore, hybrid models combining CNNs and ViTs have been developed to combine the strengths of both architectures. However, such hybrid models are difficult to interpret, which hinders their application in medical imaging. In this work, we introduce an interpretable-by-design hybrid fully convolutional CNN-Transformer architecture for retinal disease detection. Unlike widely used post-hoc saliency methods for ViTs, our approach generates faithful and localized evidence maps that directly reflect the model's decision process. We evaluated our method on two medical tasks focused on disease detection using color fundus images. Our model achieves state-of-the-art predictive performance compared to black-box and interpretable models and provides class-specific sparse evidence maps in a single forward pass.

Keywords: Self-explainability · interpretable-by-design · Hybrid CNN-Transformer · Dual-Resolution Self-Attention · Retinal fundus image.

1 Introduction

Convolutional neural networks (CNNs) are at the heart of many successful applications in medical image analysis [9], but more recently, vision transformers (ViTs) have emerged as a competitive alternative [11], demonstrating strong performance in medical imaging tasks [3, 26]. Although CNNs are highly effective at capturing complex local patterns in images, the size of their receptive field is smaller than some disease-related lesions [17]. In contrast, vision transformers leverage self-attention (SA) [27] to capture long-range dependencies, providing a more global understanding of the image. Despite these advantages, ViTs require substantial computational resources, often demanding large-scale datasets for effective training [20, 26], while also facing challenges in interpretability [16].

To address the weaknesses of both approaches, a promising alternative are hybrid CNN-Transformer architectures. Several studies have used such architectures [15,18,20,26], improving performance for tasks that require combining local features with global relationships for classification. Yet, the interpretability of such hybrid approaches has remained a challenge [18, 20, 26], as they require either techniques tailored to transformer architectures or the development of novel visualization methods [18]. To this end, either CNN-based methods have been adapted to ViTs [4, 24] or ViT-specific techniques have been proposed [1, 7, 8]. The most commonly used ViT-specific approach has been to visualize attention maps across layers, as these capture interactions between input regions. However, attention is not class-specific and merely illustrates relationships between input patches rather than their direct contribution to the model prediction [5, 16, 25]. Alternatively, post-hoc CNN-based methods like LRP [4] and GradCAM [24] have been successfully adapted to ViT by integrating gradients within the self-attention layers, offering class-wise explanations [8]. Yet, these are model-specific and struggle with hierarchical architectures like the Swin Transformer [21].

Here, we propose a novel, inherently interpretable-by-design hybrid CNN-Transformer architecture for retinal fundus image classification, combining the feature extraction strengths of CNNs with the ability of ViTs to capture long-range dependencies from dual-resolution features. Dual-resolution self-attention (DRSA) allows the model to capture both fine-grained details and global context by attending to representations at two distinct spatial resolutions. Our design integrates recent advancements such as convolutional ViTs [29], dual-resolution self-attention [15], and sparse explanations [10, 17]. We evaluated our model using two backbone CNNs—ResNet and BagNet—on two clinically relevant tasks: Diabetic Retinopathy (DR) detection and Age-Related Macular Degeneration (AMD) severity classification, using publicly available color fundus image datasets. Our hybrid model provides self-interpretability without sacrificing classification performance, challenging the myth of the accuracy-interpretability tradeoff [23]. It maintained predictive performance compared to both interpretable and non-interpretable state-of-the-art models while offering faithful explanations that accurately localize disease-related lesions—even under distribution shift—outperforming traditional post-hoc methods.

2 Developing a self-explainable hybrid CNN-ViT model

2.1 Hybrid CNN-ViT architecture

In our hybrid architecture (Fig. 1), CNN and ViT modules are used sequentially, with the output of the CNN module serving directly as the input to the transformer module. Specifically, the CNN module acted as a feature extractor, capturing local patterns. The ViT module modeled long-range dependencies between the extracted features, enhancing the model's ability to understand broader contexts. Given an input image $\mathbf{X} \in \mathbb{R}^{H \times W \times C}$—where H, and W, denote height and width, and C is the number of channels—the CNN backbone f extracts a spatial feature representation $\mathbf{Z} = f_\theta(\mathbf{X}) \in \mathbb{R}^{M \times N \times D}$, where θ denotes

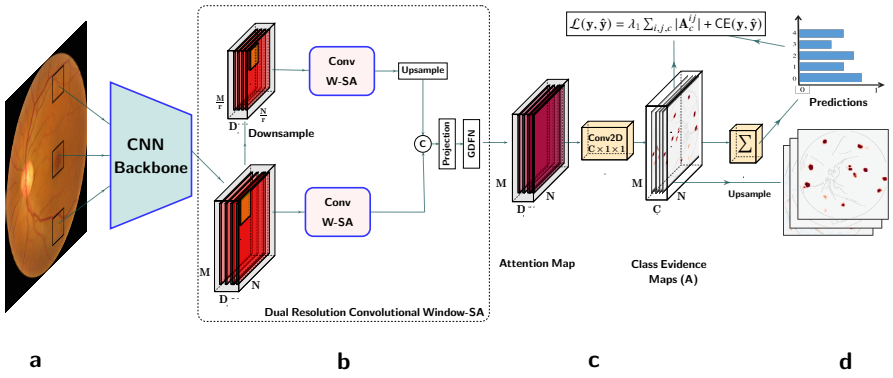

Fig. 1. Interpretable-by-design hybrid CNN-Transformer model. (**a**) Input image. The black patches illustrate the small receptive field of the CNN backbone (BagNet). (**b**) Two window-based SA modules are applied separately to the high and downsampled low-resolution feature maps, followed by feature fusion. (**c**) The high-dimensional attention map is transformed into the class evidence map $\mathbf{A}$ by applying a 1×1 convolutional classifier with C kernels, where C is the number of classes. (**d**) Spatial averaging of the class evidence maps yields predictions, whereas upsampling $\mathbf{A}$ provides explanations.

the model parameter, $M \times N$ represents the spatial size, and D is the feature dimension. We used either a ResNet50 (with a receptive field of 427×427) or a BagNet-33 (33×33) as the backbone network. Unlike the ResNet, the BagNet aggregates only local features in a bag-of-words manner [6]. The transformer module (Fig. 1b) uses a dual-convolutional window self-attention (Conv-wSA) mechanism that operates on both high- and low-resolution versions of the original feature maps to produce an attention map $\mathbf{W} = g_\phi(\mathbf{Z}_h, \mathbf{Z}_l) \in \mathbb{R}^{M \times N \times D}$. Here, $\mathbf{Z}_h = \mathbf{Z}$ denotes the high-resolution feature map, while $\mathbf{Z}_l = d(\mathbf{Z}, r) \in \mathbb{R}^{\frac{M}{r} \times \frac{N}{r} \times D}$ is the low-resolution counterpart, obtained via the downsampling function $d(.)$ with reduction factor r. The function g_ϕ, parametrized by ϕ, jointly encompasses the parameterized projection and the Gated-Dconv Feed-Forward Network (GDFN) [31]. The attention map produced by the transformer module maintains the spatial resolution of the input feature map. The classification module (Fig. 1c) comprises a convolutional layer with C kernels of size 1×1 and unit stride, producing an evidence map $\mathbf{A} = h_\psi(\mathbf{W}) \in \mathbb{R}^{M \times N \times C}$, where C represents the number of classes and ψ denotes the parameter of the classifier h. The final prediction $\hat{\mathbf{y}} \in \mathbb{R}^{1 \times C}$ is obtained by applying spatial average pooling to $\mathbf{A}$, followed by a softmax operation: $\hat{\mathbf{y}} = \text{Softmax}\big(\text{AvgPool}(\mathbf{A})\big)$. This yields a C-dimensional probability distribution representing the likelihood of each class.

2.2 Learning long-range dependencies with convolutional DRSA

To learn long-range dependencies between the convolutional features, we used a transformer module with dual-resolution self-attention (DRSA) [15], for which

a convolutional layer had replaced the linear fully connected layer (FCL) [29] as follows: $\text{SA}_h = \text{Softmax}\big(\frac{\mathbf{Q}_h \mathbf{K}_h^\top}{\alpha}\big)\mathbf{V}_h$, $\text{SA}_l = \text{Softmax}\big(\frac{\mathbf{Q}_l \mathbf{K}_l^\top}{\alpha}\big)\mathbf{V}_l$ where α is the scaling factor, and $\mathbf{Q}_h, \mathbf{K}_h, \mathbf{V}_h$ and $\mathbf{Q}_l, \mathbf{K}_l, \mathbf{V}_l$ are the queries, keys, and value embeddings generated for $\mathbf{Z}_h$ and $\mathbf{Z}_l$ using convolutional operations. The final self-attention representation is computed as: $\text{SA}_{final} = \text{GDFN}_\delta\big(\text{Proj}_\beta(\text{SA}_h + \text{Up}(\text{SA}_l))\big)$, with $\mathbf{W} = \text{SA}_{final}$, where $\text{Up}(\text{SA}_l)$ denotes the upsampled version of SA_l. This upsampled map is aggregated with SA_h and passed through a convolutional projection parametrized by β. The resulting representation is subsequently refined using a GDFN parametrized by δ, which enhances spatial structures while suppressing irrelevant features. This refinement ensures that only salient information contributes to the final predictions, thereby improving the generalization performance of the model.

2.3 Enhancing interpretability with a sparse convolutional classifier

In standard ViT and hybrid CNN-Transformer models, the classification head includes a FCL, which discards spatial information, limiting interpretability. Our architecture addressed this by preserving spatial information using convolutional operations in the self-attention module, generating attention maps that capture long-range dependencies between regions in the same window. To enhance interpretability, we replaced the FCL with a convolutional classifier, referred to as the *class evidence layer*. This layer leverages spatial information to produce class-wise evidence maps (Fig. 1), where each pixel reflects the local contribution of input regions to the final prediction. Following classification, the evidence maps are upsampled and overlaid on the input image for visualization purposes (Fig. 1d). Furthermore, the inclusion of an explicit class evidence layer enables the application of an ℓ_1 sparsity constraint on the class evidence maps $\mathbf{A}_c$, thereby enhancing interpretability [10, 17]. This leads to the following loss function:

$$\mathcal{L}(\mathbf{y}, \hat{\mathbf{y}}) = \text{CE}(\mathbf{y}, \hat{\mathbf{y}}) + \lambda \sum_{i,j,c} |\mathbf{A}_c^{ij}|. \tag{1}$$

Here, CE denotes the cross-entropy loss, and $\mathbf{y}$ represents the reference class labels. The sparsity of the evidence maps is controlled by the hyperparameter λ. The entire model is trained end-to-end using gradient descent.

3 Results

3.1 Datasets

We used two publicly available retinal fundus datasets, the Kaggle Diabetic Retinopathy (DR) [12] and the Age-Related Eye Disease Study (AREDS) [13]. The Kaggle DR dataset had 45,923 images from 28,984 subjects after applying a custom quality filtering with class distributions: 73% No DR, 15% Mild, 8% Moderate, 3% Severe, and 1% Proliferative DR. The AREDS dataset contained 34,079 images from 4,757 participants. AMD severity was grouped into

six categories [2,13]: 49%, 19%, 14%, 3%, 12%, 1% for early, moderate, advanced intermediate, early late, active neovascular, and end-stage AMD.

Images were resized to 512 × 512, normalized, and augmented with cropping, flipping, color jitter, and rotation. Datasets were split into 75% training, 10% validation, and 15% test, keeping each participant's records in the same split. We evaluated our model's ability to localize DR-related lesions against ground-truth human annotations using the IDRiD dataset [22], which provides 81 fundus images with pixel-level labels for microaneurysms (MA), hemorrhages (HE), soft exudates (SE), and hard exudates (EX). This enabled the assessment of interpretability through localization performance.

3.2 Self-explanaible hybrid models achieved SOTA performance

We first evaluated our models on multiclass DR detection and AMD severity classification. Using the ResNet50 and BagNet-33 as backbone, our model incorporated dual-resolution convolutional self-attention (DR-Conv-SA) and a GDFN module [31]. We set the reduction factor to $r = 2$, following [15], and applied max pooling. The window size was tuned ($w = 10$ for BagNet, $w = 8$ for ResNet), along with the regularization coefficient λ (Eq. 1), to balance classification accuracy and evidence map sparsity. Classification metrics are reported with 95% confidence interval (CI) lengths from bootstrapping, while inference time is reported with standard deviation (SD) over 1,000 runs on the same input.

We compared our sparse models to the dense counterparts ($\lambda = 0$), a variant using linear self-attention (SA) with fully connected layer (FCL) classifier, and several other baselines: ResNet50, BagNet33, ViT32 (input size 384), and Swin Transformer (input size 384, patch size 4, window size 12). All models were initialized with pre-trained weights from ImageNet and trained with the same setup: data augmentation, cross-entropy loss, cosine learning rate schedule, and SGD optimizer (learning rate 10^{-4}, weight decay 5×10^{-4}) on an NVIDIA A40 GPU, using PyTorch, with model selection based on the best validation accuracy.

Table 1. Classification performance on the test sets. Reported computational costs include: parameters (M), memory (MB), and average inference time (s).

	Computational Cost			AREDS AMD		Kaggle DR	
	Par.	**Mem.**	**Time**	**Acc.**	κ	**Acc.**	κ
ViT [11]	86,094	341	09.5 ± 0.1	$.76 \pm .03$	$.90 \pm .02$	$.81 \pm .02$	$.71 \pm .04$
Swin [19]	86,883	358	15.5 ± 1.1	$.78 \pm 0.2$	$\mathbf{.92 \pm .02}$	$.85 \pm .02$	$.81 \pm .03$
ResNet [14]	23,518	101	04.2 ± 0.5	$.78 \pm .03$	$.89 \pm .02$	$.85 \pm .02$	$.81 \pm .03$
BagNet [17]	16,271	193	15.1 ± 0.1	$.75 \pm .03$	$.88 \pm .02$	$.86 \pm .02$	$.83 \pm .03$
ResNet-FCL-SA	69,732	281	06.2 ± 0.2	$.78 \pm .03$	$.90 \pm .02$	$.86 \pm .02$	$.82 \pm .03$
BagNet-FCL-SA	62,501	306	27.3 ± 0.2	$.77 \pm .03$	$.89 \pm .02$	$.85 \pm .02$	$.83 \pm .03$
ResNet-Conv-SA	69,735	285	06.3 ± 0.6	$.78 \pm .03$	$.91 \pm .02$	$.85 \pm .02$	$.83 \pm .03$
BagNet-Conv-SA	62,913	310	27.3 ± 0.3	$.77 \pm .03$	$.90 \pm .02$	$\mathbf{.87 \pm .02}$	$\mathbf{.84 \pm .02}$
sResNet-Conv-SA	69,735	285	06.3 ± 0.6	$\mathbf{.79 \pm .02}$	$.90 \pm .02$	$.85 \pm .02$	$.80 \pm .03$
sBagNet-Conv-SA	62,913	310	27.3 ± 0.3	$.77 \pm .03$	$.91 \pm .02$	$.85 \pm .02$	$.81 \pm .03$

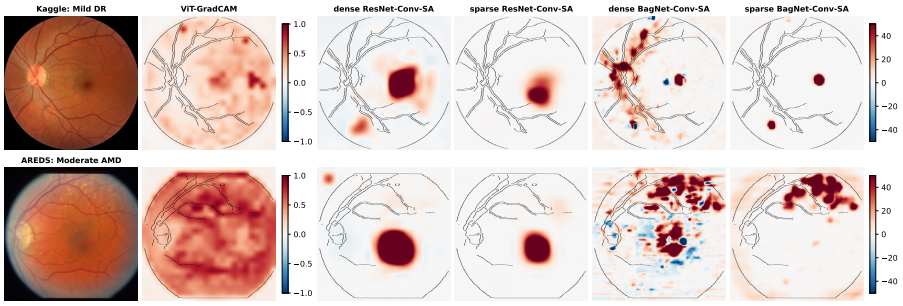

Fig. 2. Examples explanations. From left to right, heatmaps for the correctly predicted class. The first row shows an example (grade 1) from the Kaggle dataset, while the second row shows an example (grade 2) from the AREDS dataset.

Our interpretable-by-design hybrid models achieved state-of-the-art performance on both tasks. The dense model with BagNet backbone yielded the best results for DR classification, while the model with the ResNet backbone achieved the highest Cohen's kappa (κ) for AMD severity classification (Tab.1). Despite the sparsity penalty on the class activation map, the sparse models maintained competitive accuracy, with only a slight reduction in κ. Notably, for AMD detection, κ exceeded accuracy, likely due to misclassifications occurring predominantly between adjacent severity levels (Fig. 3d). Overall, the computational cost varies depending on the backbone but remains lower than that of ViTs.

3.3 Sparsity constraints enhance class evidence maps

We next compared evidence maps from our model to attribution maps generated with GradCAM [24] on the ViT baseline. As these were multiclass tasks, we only showed class evidence maps from the correctly predicted class. Our class evidence maps, obtained from the convolutional layer before average pooling, clearly highlighted input features relevant to the predicted class (Fig. 2). We noticed that GradCAM on ViT produced cluttered, hard-to-interpret heatmaps. In contrast, the hybrid ResNet-Transformer generated coarser heatmaps due to its large receptive field, while the hybrid BagNet-Transformer provided more localized explanations. The sparse models further refined this by producing sparser heatmaps, focusing decisions on smaller yet relevant retinal regions. For AMD severity classification, we observed that both the dense and sparse ResNet-Transformer models focus mainly on the macular region.

3.4 Evidence maps provide faithful and localized explanations

We quantitatively assessed the alignment of the explanations with clinical lesion-wise ground truth annotations by evaluating their precision in identifying DR lesions. Following the International Clinical Diabetic Retinopathy Scale [28],

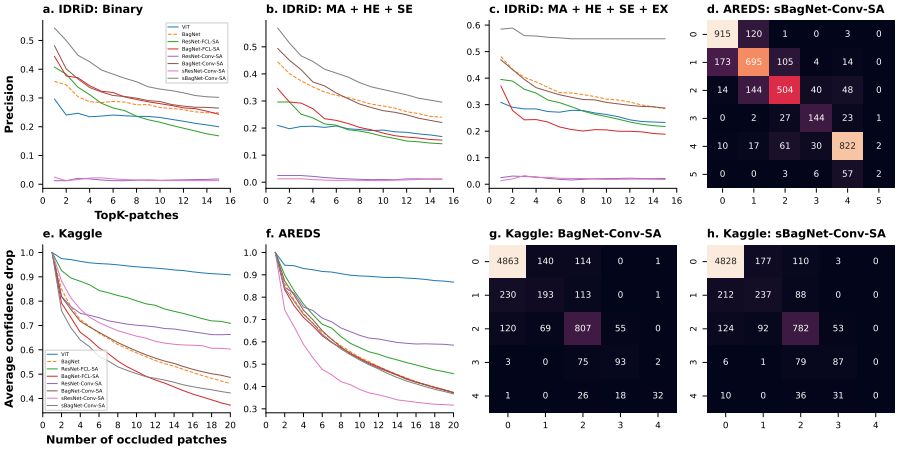

Fig. 3. Quantitative evaluation of heatmaps and confusion matrices. (**a-c**) Precision evaluation on IDRiD dataset. (**e,f**) Sensitivity analysis of different heatmaps for DR detection and AMD severity classification. (**d,g,h**) Confusion matrices of different models for DR detection and AMD severity classification on the test sets.

we evaluated three cases: (a) binary evaluation (Fig. 3a), averaging disease-class heatmaps and combining all lesion annotations; (b) severe DR (Fig. 3b), where MAs, HEs, and SEs were combined, and the precision was computed from the severe grade heatmap (c) proliferative DR (Fig. 3c), where all lesions were combined and precision was evaluated from the heatmap from the proliferative grade heatmap. Precision was measured as the proportion of positively activated regions containing lesions [17], using 33×33 non-overlapping patches to match BagNet's receptive field. For ViT and hybrid FCL models, GradCAM-generated heatmaps were used. Patches were extracted from positively activated regions. In all cases, the sparse BagNet-Transformer showed considerably higher precision than all other models and outperformed the base BagNet, suggesting that incorporating attention improved both classification and interpretability. The ResNet-Transformer with an explicit class-evidence layer performed worse, likely due to its larger receptive field producing coarser localizations (Fig. 2).

Subsequently, we additionally measured the faithfulness of the explanations by evaluating their ability to identify relevant regions for classification [30]. Using correctly classified test images, we progressively removed top-ranked patches highlighted in the heatmap and measured the resulting drop in class confidence. For DR detection, the sparse BagNet-Transformer performed best, while standard ViTs performed worst, followed by the ResNet-Transformer (Fig. 3e). In contrast, for AMD severity classification, the hybrid sparse ResNet outperformed the sparse BagNet-Transformer (Fig. 3f), likely due to the larger lesion sizes in AMD, which favor CNNs with larger receptive fields. Notably, this trend was consistent with classification results, where the ResNet backbone also excelled.

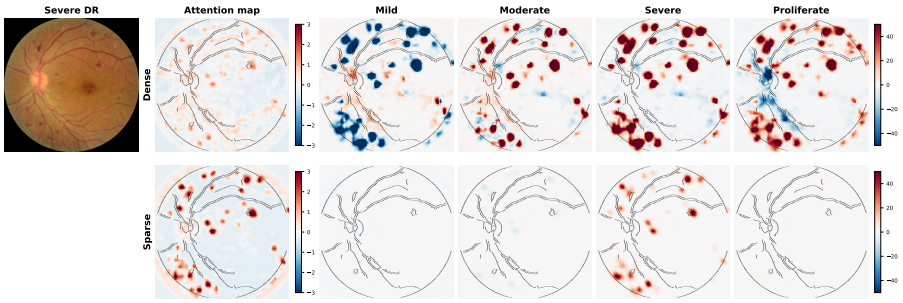

Fig. 4. Examples of multi-class explanations. Class-specific heatmaps for a Severe DR example from the Kaggle dataset. The first row displays the attention map and corresponding heatmaps from the dense hybrid model with the BagNet backbone, while the second row shows the attention map and heatmaps from its sparse version.

3.5 Our model enhances interpretability for multi-class tasks

Finally, we visualized class-specific explanations for the dense and sparse BagNet-Transformer. For DR prediction on the Kaggle dataset, both models correctly classified the example (Fig. 4). For our hybrid model, heatmaps and class probability distributions were generated in a single forward pass, with the sparse model producing more focused and localized explanations aligned with both the predicted class and clinical ground truth. In contrast, post-hoc explanations required multiple forward passes, increasing the overall inference cost. In other classes, the sparse model showed almost no positive activations, unlike the dense model, which presented a mix of positive and negative evidence. Interestingly, we observed a strong correlation between attention maps and predicted evidence maps, particularly in the sparse model. This suggests that the model effectively captures long-range dependencies in an interpretable way.

4 Discussion and Conclusion

We introduced the first inherently interpretable hybrid CNN-transformer architecture for medical image classification[3], applied to DR detection and AMD severity classification from retinal fundus images. The approach is backbone-agnostic, allowing backbone selection to be guided by disease-specific prior. We evaluated the model with two CNN backbones: ResNet50, which captured global spatial relationships relevant to AMD, and BagNet, which aggregates small local features important for DR detection. The latter is particularly noteworthy, as the SA mechanism helps overcome BagNet's limited receptive field. Conversely, ResNet's larger receptive field is better suited for AMD, which involves larger lesions. In both cases, SA enhances the model's focus on the most relevant features. Our transformer module employs dual-resolution convolutional SA to

[3] Code at `https://github.com/kdjoumessi/Self-Explainable-CNN-Transformer`

capture both global and fine-grained features while preserving strong local inductive biases. Unlike standard models with FCL classifiers, our model includes an explicit class evidence layer that produces spatial class-evidence heatmaps, enabling direct interpretability without post-hoc methods.

Interestingly, the interpretability–accuracy trade-off was relatively small, challenging the myth of the accuracy-interpretability tradeoff in self-explainable models [23]. All evaluated models achieved comparable performance, with high balanced accuracy and κ. Notably, the sparse BagNet-Transformer produced the most informative explanations for DR detection, while the sparse ResNet-Transformer yielded the best explanations for AMD severity classification.

Preliminary experiments showed that multi-head SA increased training time without improving classification, while multi-scale resolution had limited impacts and further increased both the training and inference time—particularly with the BagNet backbone (Tab. 1), due to its larger feature maps and the resulting higher SA computation cost. Following [17], we also observed that higher sparsity often led to missed detection of late-stage DR, likely due to their underrepresentation in the training set (Fig. 4h). However, our hybrid architecture mitigated this issue more effectively, demonstrating robustness in low-data settings. Overall, our findings underscore hybrid CNN-Transformer models as a strong alternative to post-hoc ViT explanations, particularly for medical imaging.

Acknowledgments. This project was supported by the Hertie Foundation, the German Science Foundation (Excellence Cluster EXC 2064 "Machine Learning—New Perspectives for Science", project number 390727645). The authors thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting KD. PB is a member of the Else-Kröner-Kolleg "ClinBrAIn".

Disclosure of Interests. The authors declare no competing interests.

References

1. Abnar, S., Zuidema, W.: Quantifying attention flow in transformers. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 4190–4197 (Jul 2020)
2. Al-Zamil, W.M., Yassin, S.A.: Recent developments in age-related macular degeneration: a review. Clinical interventions in aging pp. 1313–1330 (2017)
3. Azad, R., Kazerouni, A., Heidari, M., Aghdam, E.K., Molaei, A., Jia, Y., Jose, A., Roy, R., Merhof, D.: Advances in medical image analysis with vision transformers: a comprehensive review. Medical Image Analysis **91**, 103000 (2024)
4. Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.R., Samek, W.: On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PloS one **10**(7), e0130140 (2015)
5. Bibal, A., Cardon, R., Alfter, D., Wilkens, R., Wang, X., François, T., Watrin, P.: Is attention explanation? an introduction to the debate. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 3889–3900 (2022)

6. Brendel, W., Bethge, M.: Approximating cnns with bag-of-local-features models works surprisingly well on imagenet. International Conference on Learning Representations (2019)
7. Chefer, H., Gur, S., Wolf, L.: Generic attention-model explainability for interpreting bi-modal and encoder-decoder transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 397–406 (2021)
8. Chefer, H., Gur, S., Wolf, L.: Transformer interpretability beyond attention visualization. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 782–791 (2021)
9. Chen, C., Isa, N.A.M., Liu, X.: A review of convolutional neural network based methods for medical image classification. Computers in Biology and Medicine **185**, 109507 (2025)
10. Djoumessi, K., Berens, P.: Soft-cam: Making black box models self-explainable for high-stakes decisions. arXiv preprint arXiv:2505.17748 (2025)
11. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: 9th International Conference on Learning Representations, ICLR (2021)
12. Dugas, E., Jared, J., Cukierski, W.: Diabetic retinopathy detection (2015), `https://kaggle.com/competitions/diabetic-retinopathy-detection`
13. Group, A.R.E.D.S.R.: The age-related eye disease study (areds): Design implications areds report no. 1. Controlled clinical trials **20**(6), 573–600 (1999)
14. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
15. Ilyas, Z., Saleem, A., Suter, D., Schousboe, J.T., Leslie, W.D., Lewis, J.R., Gilani, S.Z.: A hybrid cnn-transformer feature pyramid network for granular abdominal aortic calcification detection from dxa images. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 14–25. Springer (2024)
16. Kashefi, R., Barekatain, L., Sabokrou, M., Aghaeipoor, F.: Explainability of vision transformers: A comprehensive review and new perspectives. arXiv preprint arXiv:2311.06786 (2023)
17. Kerol, D., Ilanchezian, I., Kühlewein, L., Faber, H., Baumgartner, C.F., Bah, B., Berens, P., Koch, L.M.: Sparse activations for interpretable disease grading. In: Medical Imaging with Deep Learning (2023)
18. Kim, J.W., Khan, A.U., Banerjee, I.: Systematic review of hybrid vision transformer architectures for radiological image analysis. Journal of Imaging Informatics in Medicine pp. 1–15 (2025)
19. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 10012–10022 (2021)
20. Maurício, J., Domingues, I., Bernardino, J.: Comparing vision transformers and convolutional neural networks for image classification: A literature review. Applied Sciences **13**(9), 5521 (2023)
21. Nguyen, H.C., Lee, H., Kim, J.: Inspecting explainability of transformer models with additional statistical information. arXiv preprint arXiv:2311.11378 (2023)
22. Porwal, P., Pachade, S., Kamble, R., Kokare, M., Deshmukh, G., Sahasrabuddhe, V., Meriaudeau, F.: Indian diabetic retinopathy image dataset (idrid): a database for diabetic retinopathy screening research. Data **3**(3), 25 (2018)

23. Rudin, C.: Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature machine intelligence **1**(5), 206–215 (2019)
24. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE international conference on computer vision. pp. 618–626 (2017)
25. Stassin, S., Corduant, V., Mahmoudi, S.A., Siebert, X.: Explainability and evaluation of vision transformers: An in-depth experimental study. Electronics **13**(1), 175 (2023)
26. Takahashi, S., Sakaguchi, Y., Kouno, N., Takasawa, K., Ishizu, K., Akagi, Y., Aoyama, R., Teraya, N., Bolatkan, A., Shinkai, N., et al.: Comparison of vision transformers and convolutional neural networks in medical image analysis: A systematic review. Journal of Medical Systems **48**(1), 1–22 (2024)
27. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L.u., Polosukhin, I.: Attention is all you need. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neural Information Processing Systems. vol. 30 (2017)
28. Wilkinson, C.P., Ferris III, F.L., Klein, R.E., Lee, P.P., Agardh, C.D., Davis, M., Dills, D., Kampik, A., Pararajasegaram, R., Verdaguer, J.T., et al.: Proposed international clinical diabetic retinopathy and diabetic macular edema disease severity scales. Ophthalmology **110**(9), 1677–1682 (2003)
29. Wu, H., Xiao, B., Codella, N., Liu, M., Dai, X., Yuan, L., Zhang, L.: Cvt: Introducing convolutions to vision transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 22–31 (2021)
30. Yeh, C.K., Hsieh, C.Y., Suggala, A., Inouye, D.I., Ravikumar, P.K.: On the (in) fidelity and sensitivity of explanations. Advances in neural information processing systems **32** (2019)
31. Zamir, S.W., Arora, A., Khan, S., Hayat, M., Khan, F.S., Yang, M.H.: Restormer: Efficient transformer for high-resolution image restoration. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5728–5739 (2022)

B Supplementary materials

The following publications are accompanied by supplementary materials that, while not fully included in the main papers, are available through the respective publishers.

1. **Kerol Djoumessi**, Ziwei Huang, Laura Kühlewein, Annekatrin Rickmann, Natalia Simon, Lisa Koch, and Philipp Berens. "An inherently interpretable AI model improves screening speed and accuracy for early diabetic retinopathy". *PLOS Digital Health*, 2025.

2. **Kerol Djoumessi**, Bubacarr Bah, Laura Kühlewein, Philipp Berens, and Lisa Koch. "This actually looks like that: Proto-bagnets for local and global interpretability-by-design". In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 718–728, 2024.

Note that the first paper includes textual descriptions of the supplementary materials within the main papers, while the corresponding figures and tables are provided online alongside the article on the publisher webpage[7].

[7] https://journals.plos.org/digitalhealth/article?id=10.1371/journal.pdig.0000831

This actually looks like that: Proto-BagNets for local and global interpretability-by-design

Kerol Djoumessi[1,2(✉)], Bubacarr Bah[3], Laura Kühlewein[4], Philipp Berens[1,2(✉)], and Lisa Koch[1,2(✉)]

Supplementary material

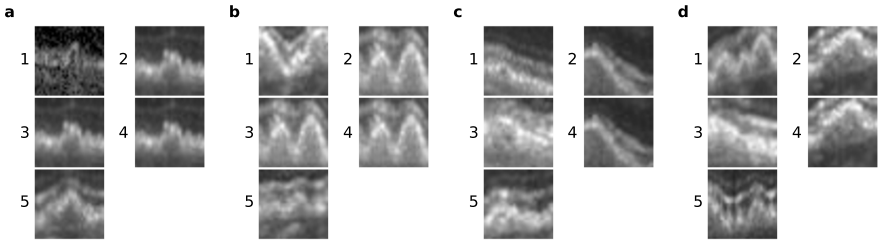

Fig. 1. Examples of some learned prototypes without adding the dissimilarity loss to prevent the model from learning redundant prototypes. **(a,b)** Prototypes 2,3, and 4 are duplicated. **(c,d)** Prototypes 2 and 4 are duplicated.

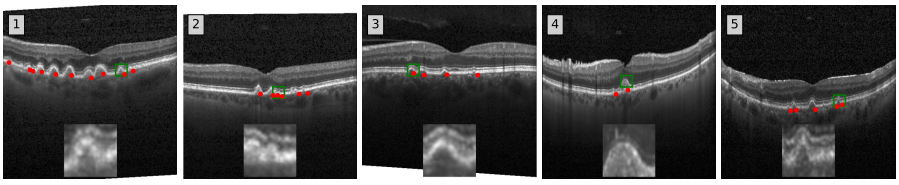

Fig. 2. Annotated training images from which the disease prototypes were extracted. The green boxes indicate the region where the learned prototypes were extracted, which are enlarged at the bottom. The red markers denote the reference annotations of drusen lesions. The number at the top indicates the prototype ID. For prototype 4, the bounding box is slightly above the lesion, probably due to a mistake when clicking on the lesion.

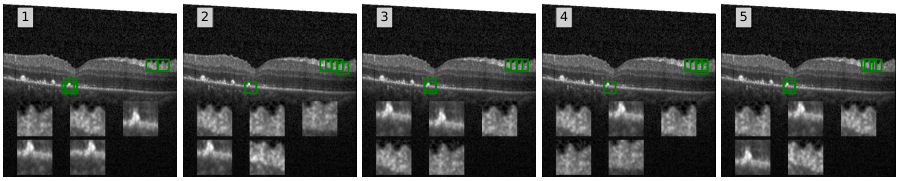

Fig. 3. Example of suspicious regions highlighted by Proto-BagNet on a disease image where the ophthalmologist did not find drusen lesions. The region highlighted near the Retinal Pigment Epithelium are sub-retinal deposits which are not typical drusen lesions but wringing of the ganglion cell layer.

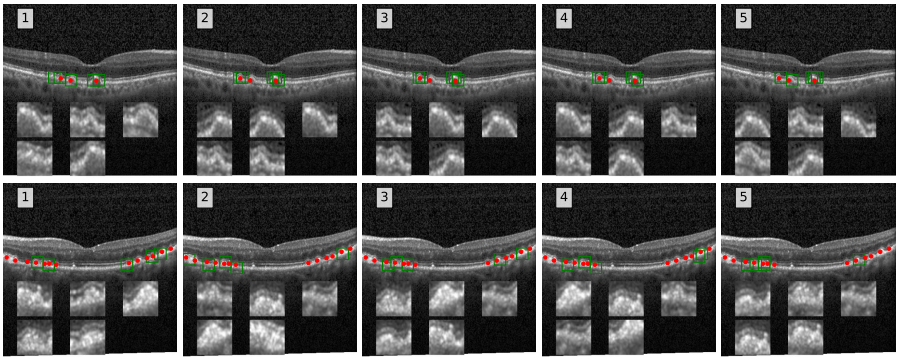

Fig. 4. Two examples of suspicious lesions extracted from each prototype similarity map on disease images. Drusen (annotated with red markers) are detected with high precision.

Table 1. Classification performance with confidence intervals (CIs) for drusen detection on validation and test sets. CIs are derived from bootstrapping with n=1000.

	Validation set				Test set			
	Accuracy	AUC	Recall	Precision	Accuracy	AUC	Recall	Precision
ResNet-50	.99 ± 1e-4	.99 ± 1e-4	.98 ± 2e-4	.99 ± 2e-4	.99 ± 1e-4	.99 ± 1e-4	.98 ± 2e-4	.99 ± 2e-4
dense BagNet	.99 ± 1e-4	.99 ± 1e-4	.98 ± 2e-4	.98 ± 2e-4	.99 ± 1e-4	.99 ± 1e-4	.98 ± 2e-4	.98 ± 2e-4
ProtoPNet	.99 ± 1e-4	.99 ± 1e-4	.98 ± 2e-4	.97 ± 2e-4	.99 ± 1e-4	.99 ± 1e-4	.98 ± 1e-4	.97 ± 1e-4
Proto-BagNet	.98 ± 1e-4	.99 ± 1e-4	.94 ± 3e-4	.98 ± 2e-4	.98 ± 1e-4	.99 ± 1e-4	094 ± 3e-4	.98 ± 2e-4

C Related contributions

I contributed to the following works, which are not part of this thesis.

1. Julius Gervelmeyer, Sarah Müller, **Kerol Djoumessi**, David Merle, Simon J Clark, Lisa Koch, Philipp Berens. "Interpretable-by-design deep survival analysis for disease progression modeling". In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 718–728, 2024.

2. Samuel Ofosu Mensah, **Kerol Djoumessi**, Philipp Berens. "Prototype-Guided and Lightweight Adapters for Inherent Interpretation and Generalisation in Federated Learning". In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 464-473, 2025.

3. Samuel Ofosu Mensah, Jonas Neubauer, Murat Seçkin Ayhan, **Kerol Djoumessi**, Lisa Koch, Mehmet Murat Uzel, Faik Gelisken, Philipp Berens. "Clinically Interpretable Deep Learning via Sparse BagNets for Epiretinal Membrane and Related Pathology Detection". *medRxiv preprint*, 2025.