Hardware-Aware Machine Learning Methods for Medical Edge Devices

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

vorgelegt von
Julia Helga Werner
aus Bad Soden am Taunus

Tübingen
2026

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

Tag der mündlichen Qualifikation:	20.05.2026
Dekan:	Prof. Dr. Thilo Stehle
1. Berichterstatter:	Prof. Dr. Oliver Bringmann
2. Berichterstatter:	Prof. Dr. Jochen Hampe
3. Berichterstatter:	Prof. Dr. Thomas Küstner

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

Abstract

Faculty of Science

Chair for Embedded Systems

Hardware-Aware Machine Learning Methods for Medical Edge Devices

by Julia Helga Werner

Real-time embedded edge devices are of great importance in many applications, such as autonomous driving, robots or smartwatches. Additionally, various medical procedures involve embedded, small in-body sensor edge devices. Equipping such devices with artificial intelligence can improve the procedures by incorporating new functionalities. However, this is often accompanied by a high demand for energy and computational resources if the models are not optimized accordingly. For some applications involving in-body edge devices equipped with machine learning models, the neural network parameters must be stored directly on-device and the model executed locally. Notably, resource-constrained devices impose stringent requirements on machine learning models in respect to on-chip area and electrical energy consumption. These restrictions need to be considered in the final model design. Deep learning methods involving neural networks with millions or even billions of parameters or operations that cannot simply be transferred to hardware, are not a viable solution. Furthermore, there is a necessity of using lightweight, quantized models in fixed-point representation to realize efficient inference on hardware. Storing model parameters in lower precision potentially impairs the overall performance of the classifier, which needs to be addressed by dedicated techniques, such as hardware-aware training. Additionally, potential challenges involve general data sparsity and class imbalances, which often occur in medical datasets since pathologies are naturally underrepresented compared to healthy samples. Importantly, if well-designed, machine-learning-based decision models provide new energy-saving functionalities that can lower the energy demand of the whole system. This thesis specifically addresses these problems by examining two important medical applications: the Video Capsule Endoscopy, a methodology to investigate the otherwise inaccessible small intestine using a small, pill-sized capsule and seizure detection using neuroimplants intended for drug-resistant epilepsy patients.

The main objective of this thesis is to overcome the described challenges and design artificial intelligence-based classification models suitable for tiny edge devices as present in both introduced medical applications. It is further expected that other medical applications can benefit from the presented methods as well. Overall, this work is dedicated to the development of hardware-aware, specialized machine learning techniques for the Video Capsule Endoscopy and preictal seizure detection. The approaches are tailored for an on-device application, providing the groundwork for future innovations and enhancements, such as an actively controlled capsule. For both applications, hybrid models are proposed, combining machine learning classifiers based on deep neural networks with time-series techniques, such as Hidden Markov Models, to solve these challenges. The resulting methods are accurate, highly efficient and are verified on FPGA-based hardware demonstrators to measure their power consumption. This enhances both medical procedures involving low-power edge devices without increasing the energy demand of the whole system.

Zusammenfassung

Eingebettete und in Echtzeit fungierende Edge Geräte sind heutzutage von großer Wichtigkeit in vielen Bereichen, wie zum Beispiel beim autonomen Fahren, in Robotern oder in Fitnessuhren. Hierbei stellt die Medizin mit ihren vielschichtigen Anwendungsgebieten keine Ausnahme dar, wobei insbesondere auf Sensoren basierende Edge Geräte direkt innerhalb des Körpers zur Anwendung kommen. Die Nutzung solcher Edge Geräte in Kombination mit künstlicher Intelligenz kann dabei Verfahren wesentlich verbessern, beispielsweise durch die Bereitstellung neuer Funktionen, die in Echtzeit und lokal sinnvolle Ergänzungen bieten. Wenn integrierte, maschinelle Lernverfahren jedoch nicht hinsichtlich der speziellen Anforderungen solcher kleinen Geräte optimiert sind, ist dies oft mit einem hohen Ressourcen- und Energiebedarf verbunden. Während der Entwicklung solcher spezialisierten Verfahren ist es daher essentiell, dass die Restriktionen dieser an Ressourcen limitierten Geräte berücksichtigt werden. Neuronale Netze mit vielen Millionen Parametern können nicht direkt auf die Hardware transferiert werden und sind daher oft nicht umsetzungsfähig. Des Weiteren fordert die Inferenz in der Hardware die Umsetzung der Modelle in Festkommaarithmetik, sodass die Ausführung neuronaler Netze zum Beispiel mittels eines Hardwarebeschleunigers realisiert werden kann. Die Modellparameter in geringerer Präzision zu speichern beeinträchtigt jedoch tendenziell die Klassifizierungsgenauigkeit. Dies sollte berücksichtigt werden, indem die Hardware während des Trainings der Modelle mit einbezogen wird. Sofern Modelle, die auf maschinellen Lernverfahren basieren, sinnvoll konzipiert werden, können sie energiesparende Funktionen bereitstellen und somit den Energiebedarf des gesamten Systems sogar verringern. Zusätzlich bestehen weitere Schwierigkeiten, wie stark ausgeprägte Klassenunausgeglichenheiten und das Vorhandensein geringer Mengen an Daten mit positiven Befunden in medizinischen Datensätzen, in denen Daten von gesunden Personen typischerweise in größerer Menge vorhanden sind als Pathologien. Die vorliegende Arbeit beschäftigt sich mit zwei wichtigen medizinischen Anwendungen: Die Videokapselendoskopie, eine minimal-invasive Untersuchungsmethode, bei der der Dünndarm auf verschiedene Erkrankungen untersucht wird, der ansonsten durch gegenwärtige Standardverfahren im Gegensatz zu den übrigen Organen des Verdauungstrakts unzugänglich ist. Das zweite Anwendungsthema fokussiert sich auf die Anfallsdetektion mittels Neuroimplantaten, die für Epilepsiepatienten intendiert sind, die weder durch Operationen behandelt werden können, noch erfolgreich auf Medikamente ansprechen.

Das Ziel der vorliegenden Arbeit ist es, die identifizierten Schwierigkeiten gezielt zu überwinden und auf künstlicher Intelligenz basierende Klassifizierungsmodelle zu konzipieren, die für kleine Edge Geräte geeignet sind, wie sie in den beiden vorgestellten Anwendungen vorkommen. Gleichzeitig ist zu erwarten, dass auch andere medizinische Anwendungen hiervon unterstützt werden können. Insgesamt befasst sich die vorliegende Arbeit mit der Entwicklung von hardware-effizienten und spezialisierten maschinellen Lernverfahren, sodass die vorgestellten Anwendungen von diesen Methoden profitieren können. Die bereitgestellten Modelle werden dabei für die direkte Anwendung auf den Geräten zugeschnitten und fungieren als fundamentale Ausgangspunkte für zukünftige Verbesserungen und Innovationen. Für die vorgestellten medizinischen Anwendungen werden daher Klassifizierungsmodelle entwickelt, die auf neuronalen Netzwerken basieren und zusätzlich mit Techniken aus der Zeitreihenanalyse kombiniert werden, wie Hidden-Markov-Modelle und Viterbi-Dekodierung. Die entwickelten Methoden sind sowohl akkurat als auch effizient. Sie werden abschließend auf FPGA-basierten Hardware-Demonstratoren evaluiert, um Rückschlüsse auf den Energiebedarf zu ziehen. Dadurch können medizinische Verfahren, die auf eingebetteten Edge Geräten basieren, maßgeblich verbessert, sowie der Energieverbrauch deutlich gesenkt werden.

Danksagung

Ich bedanke mich ganz besonders bei Professor Oliver Bringmann für die Möglichkeit, meine Promotion an seinem Lehrstuhl unter seiner Betreuung durchzuführen. Er hat sich stets Zeit für mein Forschungsprojekt genommen und durch die bereichernden Gespräche meinem Promotionsprojekt wichtige Impulse verliehen. Vor allem wird mir die wertschätzende and angenehme Arbeitsatmosphäre in Erinnerung bleiben.

Weiterhin möchte ich mich ausdrücklich bei Professor Jochen Hampe für die Unterstützung meines Promotionsvorhabens, sowie die für diese Arbeit essentielle Bereitstellung von Videokapselendoskopie Studien und seine medizinische Expertise bedanken. Ohne diese motivierende Unterstützung wäre meine Arbeit so nicht möglich gewesen.

Zudem danke ich Professor Thomas Küstner für die interessanten Diskussionen, neuen Anregungen und die Begutachtung dieser Arbeit.

Außerdem möchte ich meinen Kollegen für den insbesondere fachlich spannenden Austausch danken.

Ein großer Dank gilt meinem Vater und meinen Geschwistern, die mich in jeder Hinsicht immer unterstützt haben.

Besonders möchte ich mich bei meinem Partner für seinen bedingungslosen Rückhalt und seine Unterstützung während meiner gesamten Promotionszeit bedanken.

Contents

Chapter 1

Introduction and Motivation

Many essential medical applications involve embedded edge devices, which facilitate new diagnostic and therapeutic possibilities. Edge devices encompass a variety of different systems, from cloud-based solutions to in-body sensor edge devices. However, for many medical applications, cloud-based solutions are unsuitable, e.g. due to the need for a consistent broad-band and non-disruptive network connection or data privacy. This highlights the need for medical edge devices, which allow the processing of medical data directly on-site.

In-body sensor edge devices are of main relevance in various medical applications, and are the primary focus in this work. Equipping such devices with Machine Learning (ML) methods can provide useful features, such as enhancing diagnostics or providing energy-saving functionalities. If ML models are integrated into sensor edge devices, starting with the model selection, one needs to consider the limitations of the devices. In-body sensor edge devices are typically very small and therefore have only limited power and resources, leading to the accompanied risk of early battery depletion before the procedure is completed. Thus, storing neural network parameters on-device and executing the model locally demands careful consideration of both the target edge device and the utilized ML model. In this context, application-specific demands regarding latency, available on-chip memory and the power consumption need to be included in the hardware design. Importantly, while adequate hardware accelerators can be precisely tailored for each application, leading to a minimal power consumption, it is desirable to find a reasonable trade-off between the flexibility which allows deployment and functionality for a wide range of models, as well as the total area cost and energy demand. Besides adapting the hardware accordingly, designing the ML model in a hardware-aware manner is another reasonable and important aspect. In this work, the latter is addressed through the example of two important real-world applications, the Video Capsule Endoscopy (VCE) for a detailed investigation of the small intestine and seizure detection of epilepsy patients with neuroimplants. Potential solutions can be offered by equipping such sensor edge devices with hardware-aware Artificial Intelligence (AI), that can not only prolong the battery life by enabling smart decisions, but can additionally provide important features, which help to improve such medical procedures.

This work proposes hardware-aware machine learning methods tailored for two medical applications, that involve in-body sensor edge devices, and provides the first crucial step by designing and testing adequate hardware-aware ML methods. In this work, ML models are considered hardware-aware, if they comply with the constraints imposed by the medical applications involving low-power edge devices. In this context, the limitations are mainly addressed by restricting the model complexity, employing fixed-point representation and considering only hardware suitable operations in neural networks. Ultimately, for some models the hardware suitability is proven by hardware simulation. Addressed in this

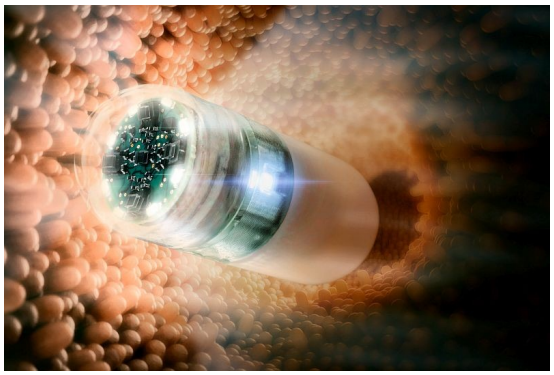
thesis, the first application is the VCE for the inspection of the Gastrointestinal (GI) tract, the second one is the seizure detection with a neuroimplant for epilepsy patients. In the following, both applications are introduced along with possible challenges, followed by outlining the contributions of this thesis.

Video Capsule Endoscopy

In the early 90s, the idea of a capsule endoscopy was initially formed by Gavriel J. Iddan and Paul Swain [IS04; Idd+00], who independently developed the concept of a wireless capsule endoscopy and later worked jointly in this matter. This minimally invasive medical procedure involves a pill-sized capsule, which mainly consists of an optical dome, a lens, multiple white Light-Emitting Diodes (LED), a camera, a battery and an antenna. This structure is depicted in Figure 1.1a for the first capsule endoscopy model M2A™ (standing for "Mouth to Anus") from Given Imaging Ltd., which was later renamed to PillCam Small Bowel (SB) [SST16]. In the beginning, this involved Charge-Coupled Devices (CCD) image sensors that operated in theory only for 10 minutes before battery depletion. However, this was later substantially advanced by the use of Complementary Metal Oxide Semiconductor (CMOS) imaging sensors, leading to images of higher quality and a lower power consumption at the same time. The first capsule was then swallowed in 1999 by Paul Swain himself [IS04], with an initial battery runtime of only 2 hours.

Eventually, in the early 2000s, the capsule endoscopy was approved by the Food and Drug Administration (FDA). From then on, this diagnostic technique was used to capture images and to analyze the GI tract for potential diseases, such as cancer, inflammatory and immune diseases, generating up to 60,000 images per patient [Fir+03; Pen+04; Sai+20]. The VCE is typically indicated when an esophagogastroduodenoscopy and colonoscopy did not lead to any substantial findings and the diagnosis remains open or requires additional information. After swallowing the device, it traverses forward by the natural peristalsis of the digestive tract and while it moves along, captures images with an adaptive or defined frame rate (see Figure 1.1b).

(A) M2A™ from Given Imaging Ltd. with an optical dome (1), lens holder (2), lens (3), LEDs (4), CMOS imager (5), battery (6), Application-Specific Integrated Circuit transmitter (7) and antenna (8) from [Yu02]



(B) VCE in the small intestine of the digestive tract from [IZM]

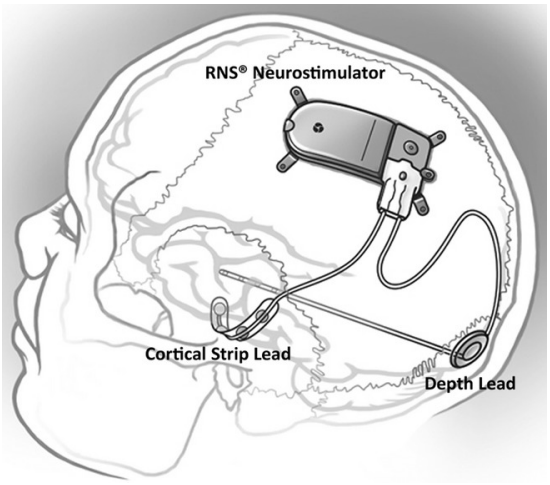
Importantly, in contrast to a classical endoscopy, the VCE allows for scanning the large middle fraction of the GI tract, the small intestine, which otherwise remains unknown.

Due to its limited size, e.g. the PillCam™ SB 3 from Medtronic, often considered as a gold standard, has a width of 2.62 cm and a diameter of 1.14 cm [Med], the number of sensors and additional features, that can be added, are also limited and further restricts the on-device storage possibility of images. While there are VCE capsules available, that store captured images on-device directly [Zwi+19], this is accompanied by drawbacks such as required collection of the capsule after the procedure, and no option for real-time assessment or decision-making. Devices, such as the PillCam™ SB 3, transmit captured images directly to a computer for further medical assessment. During this procedure, it is essential to cover the relevant part of the digestive system before the capsule's battery is depleted. While the PillCam™ SB 3's operating time is listed with ≥ 8 h, the PillCam® SB2-ex for example allows 12 h of battery time [Mon+16]. However, for some patients the complete traversal of the capsule might exceed 8 or even 12 h, potentially leaving parts of the relevant regions not fully revealed [SST16]. Thus, implementing techniques, that prolong the battery lifetime and simultaneously add more functionalities, will enhance the VCE further. Other future innovations include adaptive sampling rates and an active control of the capsule, including zooming in on critical regions with anomalous indication.

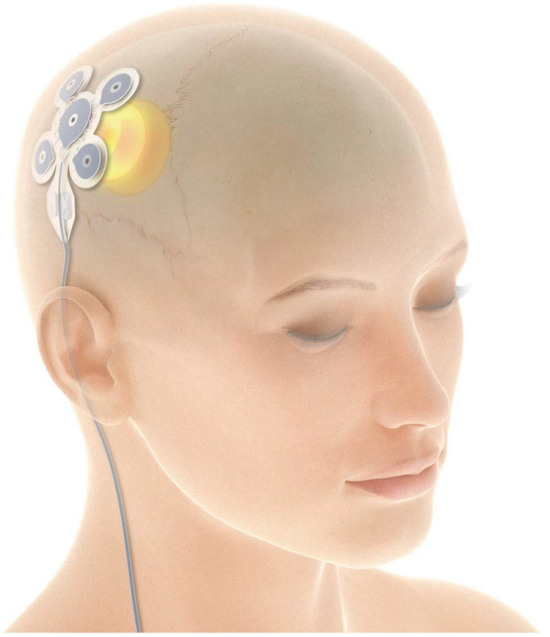
Epilepsy and Seizure Detection

Epilepsy is one of the most prevalent chronic, neurological diseases worldwide, with over 70 million people affected by this condition [Thi+19]. One crucial symptom of this disease is reoccurring seizures, originating from an excessive electrical activity, primarily of cortical neurons that can result in brief mental lapses or even unconsciousness [Fis+14]. The seizures can be further categorized into three groups depending on its onset region: focal, generalized and unknown [Thi+19]. Approximately in 60% of cases, focal seizures are classified, meaning that the seizures affect only one hemisphere. While there are anti-epileptic drug options available to reduce such seizures, about one third of patients cannot be considered for this kind of treatment or do not respond to this medication [WHO; Dun+06]. In this case, surgery might be conducted to remove the respective brain region, where the seizure originates, leading to the clearance of seizures in $50\% - 80\%$ of patients [RCR14; De +11; Mat+17]. If neither surgery nor anti-epileptic drug treatment relieves the symptoms, an alternative option is provided by novel neurostimulation techniques.

Such medical devices stimulate the focus brain area electrically resulting in a balancing of otherwise increased brain activity. The stimulation is either strictly scheduled in an open-loop manner or only initiated after a seizure is actually recognized, in a closed-loop manner. This alternative treatment option of neurostimulation devices can be categorized into three groups: Vagus Nerve Stimulation (VNS), Deep Brain Stimulation (DBS) and Responsive Stimulation (RNS). VNS devices were the first neurostimulators, that were approved by the FDA in 1997 and can be used for patients with refractory epilepsy [Lan02; Ben02; Mor+13]. It is based on electric stimulation of the vagus nerve in the neck up to the brain, which is attached to a generator device implanted in the chest. Deep brain stimulation describes the stimulation by electrodes, which are implanted directly within the brain [Loz+19]. Responsive stimulation is only applied upon detection of a seizure, based on variations within the frequency or amplitude of Electroencephalography (EEG) data [SMW08; Pet+01]. Thus, besides the capturing of the relevant data, such implants require an algorithm or model, e.g. a neural network, which can perform seizure detection in real-time. In this work, only the third group is closer investigated, for which we focus on the efficient on-device seizure detection. One product that provides neurostimulation upon seizure detection is the RNS NeuroPace neurostimulator, which consists of a stimulator, a cortical strip lead and depth lead with four electrodes each implanted nearby the seizure foci, as illustrated in Figure 1.2a.



(A) RNS® NeuroPace Neurostimulator with depth and cortical strip lead, Figure reproduced from [Hec+14] (under CC BY-NC-ND license).



(B) Transcranial neurostimulation with a neuroimplant for epilepsy patients (EASEE® System) from [KS20].

Additionally, the system requires a patient data transmitter for monitoring as well as a telemetry wand for wireless communication [SMW08]. They employ area [Ech+99], line-length [Est+01] and half-wave [Got82] algorithms for seizure detection based on the captured EEG data. Another example of a neurostimulator is the EASEE® system [Sch+23] from Precisis [Pre] for refractory focal epilepsy. This is implanted subgaleal at the individual onset seizure regions and equipped with a pulse generator and electrodes for stimulation [Sch+23], as shown in Figure 1.2b. This system is less invasive than NeuroPace, since it is implanted epicranial and the electrodes do not actually touch the brain. In the future, it could be extended with a Convolutional Neural Network (CNN) to a responsive stimulation system upon detecting voltage fluctuations, allowing for real-time seizure detection on-site. However, the addition of AI models to such devices is still in the beginning and requires careful consideration of the device specifications, including very limited memory and power.

Sparsity and Challenges of Medical Data

Relying on medical data for the successful development of AI models leads to several challenges. For example, the training of neural networks requires a large amount of labeled training data; otherwise the training is prone to overfitting, and the model merely memorizes the training data without generalizing to new data. However, a large quantity of annotated medical data is often not available, which complicates the overall development process. One reason for this lies in the fact that medical data is often subject to various data privacy regulations [DP15] in most countries and thus, the available amount of data is restricted. Furthermore, medical experts are often required for the time-consuming work of annotation, for example, for labeling VCE images or EEG seizure periods, limiting the overall amount of annotated data available [Sme+21]. Additionally, this is often accompanied by inconsistencies in terms of labeling [CHP19; Ron+12], e.g. which anomaly is seen on VCE images, or the exact determination of the beginning of a seizure. Finally, for both applications, another challenge lies in the substantially higher number of normal

data compared to anomaly data [Sme+21; Sho09], leading to an overrepresentation of healthy data, which then is often reflected with a bias in the final classification performance. Another difficulty specifically occurs in the seizure detection task: a voltage deviation in the EEG data can be measured, if recorded non-invasively, compared to invasive EEG recordings with electrodes implanted subgaleal or even directly within the brain [RMK16]. Overall, it is essential to overcome these challenges of limited data availability in order to make substantial progress in this field.

Challenges of Hardware-Aware Machine Learning Methods

Aside from the difficulties that are typically encountered in medical data sets, which complicate the training of machine learning models in this context, other restrictions are introduced by the need of inference on small edge devices. Neural networks exhibit a superior performance, but also have become progressively more complex with a high computational complexity [Arn+21; Liu+21; Sze+17], which ultimately requires a larger amount of memory and results in an increased power consumption during inference. Different aspects generate limitations on the adequate ML methods for tiny edge devices. Firstly, the overall model size needs to be considered, since the storage of the model directly influences the required memory, area and energy consumption. Modern neural network architectures with many millions or even billions of parameters, such as vision transformers, easily surpass the requirements on the classification accuracy, but (currently) require far too many computational and memory resources to be implemented on edge-devices for our target applications. From a chip design point-of-view: large deep neural network architectures often require too much electric energy and area for the applications that are considered in this thesis. Necessary adjustments can be either in the form of an increased hardware architecture or by adapting the machine learning model for the target application accordingly. Within this work, the latter is being implemented. Other criteria to assess the performance of a model include the number of necessary Multiply-Accumulate (MAC) operations per model evaluation, which directly influences the power consumption. The number of MAC operations is thereby also determined by the size and structure of the network. Furthermore, ML models are typically designed in floating-point representation; however, this is very costly in hardware and low-power architectures typically employ fixed-point arithmetic [PH16; JV23]. Replacing floating-point arithmetic by fixed-point representation results in an improved energy efficiency, decreased latency as well as enhanced computational throughput [TNR00; Gup+15] but on the other hand reduces the expressivity of the neural network architecture and complicates the training of the network. Thus, even while potentially impairing the classification performance due to a lower precision, it is desirable to implement fixed-point arithmetic if targeting low-power in-body sensor edge devices. Integrating AI into edge devices is therefore a critical step where trade-offs between classification performance and model complexity need to be reconciled.

1.1 Aim and Contribution of this Thesis

Integrating machine learning methods into small, medical edge devices paves the way for innovative medical products, with increased battery life, improved capabilities and novel functionalities. Future visions include a pill-sized low-power capsule which can easily cover the whole GI tract and transmit a comprehensive evaluation of intestinal findings to the corresponding medical doctor, which then facilitates personalized and targeted therapy. With respect to the neuroimplant, one can imagine a long-lasting specialized edge device, which efficiently suppresses seizures in real-time upon accurate detection and thereby,

alleviates sufferings from epilepsy patients, who might not have other treatment options. ML models often require many billions of operations per second, however, such demands cannot be met by many current edge devices [CPC16]. Instead, tiny edge devices impose strict constraints on the employed models due to limited on-chip memory and short battery lifetime.

Both introduced applications can benefit from incorporating ML models, however, the process of selecting adequate models is accompanied by several challenges as previously mentioned. Current state-of-the-art VCE devices cannot reliably cover the whole small intestine due to a limited battery lifetime. Neural implants can benefit from small AI models by stimulating only upon seizure detection and by lowering the computational complexity of the implemented models. Besides overcoming the difficulties of data sparsity, class imbalance and potentially mislabeled data in the medical field, adequate ML models for the neuroimplant and the VCE not only need to obtain a good classification performance but must additionally conform to the given hardware constraints. This thesis aims to address the following research objectives to overcome the mentioned challenges and step towards intelligent sensor in-body edge devices for medical applications.

Research Objectives

1. Provision of an accurate hardware-aware EEG-based seizure detection pipeline targeting neuroimplants, competing with current baselines.
2. Hardware simulation of proposed seizure detection approach outperforming current baselines with low energy demand.
3. Provision of hardware-aware ML models for the localization within the GI tract during VCEs, competing with current baselines.
4. Enhancing anomaly detection in VCEs, outperforming current baselines in software.
5. Realizing joint localization and anomaly detection in the GI tract using a single model.
6. Saving energy by omitting unnecessary image preprocessing steps and model simulation on a simple VCE demonstrator.
7. Demonstrating the prolonged battery lifetime of a VCE device incorporating an AI-based decision model compared to a standard capsule without ML models integrated.

First, relevant theoretical background to this work is outlined in Chapter 2. The objectives are addressed in detail in Chapter 3 by integrating and recapitulating various peer-reviewed publications by this author and additional experiments. The full publications are attached in the Appendix A. The first two objectives include the provision of a hardware-aware EEG-based seizure detection pipeline and its sucessful hardware simulation to demonstrate its suitability for low-power neuroimplants. These objectives are discussed and fulfilled in the Publication “Energy-Efficient Seizure Detection Suitable for Low-Power Applications” (see Publication A1) by providing a hardware-suitable seizure detection pipeline using lightweight quantized CNNs and time-series analysis. This was further simulated on an ultra-low power hardware accelerator, that can efficiently execute 1-Dimensional (1D)-CNNs. The methodology and results are summarized in Section 3.1, which provides further insights into the model and channel selection. The third objective encompasses the

localization within the GI tract during VCEs. In this work, the term localization generally refers to the classification of organs in the GI tract over time rather than the execution of segmentation, as may be the standard in other fields. Such anatomical classification enables selective image transmission when the region of interest is reached, leading potentially to substantial energy savings. This is accomplished in the Publication "Precise Localization within the GI Tract by Combining Classification of CNNs and Time-Series Analysis of HMMs" (see Publication A2) by combining a lightweight CNN with a Hidden Markov Model (HMM) and Viterbi decoding to leverage the time-series properties of the VCE videos. The localization within the GI tract allows discarding images which are not of interest until the stomach is exited to potentially save energy due to omitting the transmission step of these images. This is addressed in Section 3.2.1, by describing the used approach and the most important results along with additional experiments on a quantized approach and a reduction of the input feature map. The fourth objective is addressed in the Publication "Enhanced Anomaly Detection for Capsule Endoscopy Using Ensemble Learning Strategies" (see Publication A3) by performing anomaly detection using an ensemble method of neural networks and anomaly tailored ML approaches. This is further examined in the Publication "Reliable Mislabel Detection for Video Capsule Endoscopy Data" (see Publication A5). The main objective of both manuscripts is an improved detection of pathologies during VCEs. The methodology and results are outlined and discussed in Section 3.2.2 as well as Section 3.2.4. The publication "Seeing More with Less: Video Capsule Endoscopy with multi-task Learning" (see Publication A4) complies with the fifth objective of realizing joint localization and anomaly detection within the GI tract, by presenting a multi-task model capable of VCE image-based organ detection along with anomaly detection in a single model. The experiments and results are summarized in Section 3.2.3. The two last objectives are to save electrical energy by omitting unnecessary image preprocessing steps and to demonstrate a prolonged battery lifetime of a VCE device due to an included AI-based decision model compared to capsules without any ML models integrated. This is addressed in the Publication "Smart Video Capsule Endoscopy: Raw Image-Based Localization for Enhanced GI Tract Investigation" (see Publication A6) by realizing CNN training and evaluation directly on raw Bayer pattern images as received by a miniature camera sensor in a VCE capsule. Additionally, the prolonged battery lifetime due to an AI-based decision tool is demonstrated on hardware. The methodology and the results are then discussed in Section 3.2.5, followed by additional experiments of visibility assessment based on VCE images. The work is then concluded with a summary and an outlook discussing future prospects of the given applications. Finally, the Appendix A lists all publications, which this thesis is based on in full length. The presented work builds upon current research and lays a new foundation for hardware-aware machine learning models for two important medical procedures: the VCE and seizure detection with neuroimplants. Based on this work, subsequent optimization with hardware software co-design can be conducted.

Chapter 2

Background

This chapter introduces the fundamentals of the used methodologies and concepts, such as CNNs, ensemble learning, Support Vector Machines and random forest models. This includes an introduction of key hardware metrics to assess the presented hardware simulation results. Furthermore, HMMs and Viterbi decoding are introduced, which are used for time-series analysis in this work. Next, the retrievable information from EEGs is described along with potential challenges when handling this type of data. Finally, the utilized medical datasets for both applications are outlined providing background on sample numbers, class occurrences and data imbalances.

2.1 Convolutional Neural Networks

Neural networks are nowadays an essential part of machine learning, excelling in speech recognition, text analysis and image classification. CNNs are a very commonly used subset of feedforward neural networks, which are prominently used for the classification of VCE images [Rus+21; Jai+21; Aok+19; Sme+21] as well as the EEG-based seizure detection [Zho+18; Wei+18; JSA20; Sho+21]. As feedforward neural networks, they exclude feedback loops, and consist of multiple stacked convolutional, pooling and fully connected layers, as well as different activation functions. The main building blocks are formed by the convolutional layers, which compute the convolution operation with a filter, add a bias term to the output and pass it through a nonlinear activation function. CNNs can be used to address a variety of obstacles, including but not limited to image and speech recognition [LB+95; Abd+14], image segmentation [Min+21; Jha+20], but also medical data analysis [Fri+18; Far+21]. In the following, the fundamental building blocks of the neural networks are introduced that were used throughout the thesis, starting from 2D-convolutional layers, which are the core component of CNNs for image classification tasks. The input and output of such layer is structured in several channels (e.g. different color channels at the input layer for image classification tasks). As the number of input and output channels varies, each layer has N input channels and M output channels. Restricted to a single channel, the input and output feature map are two-dimensional arrays, whose size is denoted by $D_F \times D_F$ and $D_O \times D_O$, respectively. As a simplification, we assume here that the height and width coincide for the input and output features. For the input layer of an image classification task, the value D_F corresponds to the resolution, i.e. number of pixels in each direction of the image. A 2D convolution then consequently produces an output feature map O of size $D_O \times D_O \times N$ corresponding to the height $\times$ width with N output channels for a filter K of size $W \times W \times M \times N$ with $W \times W$ being the spatial size applied to an input feature map F of size $D_F \times D_F \times M$ with M input channels [Guo+19]:

$$O_{k,l,n} = \sum_{i,j,m} K_{i,j,m,n} \cdot F_{k+i-1,l+j-1,m}. \tag{2.1}$$

Generally, the size of the output feature map (i.e. D_O) is determined by the size of the input feature map D_F and the padding applied to the input feature map. We have $D_O = D_F$, if sufficiently many zeros are added to correct for the effect of the width of the filter in the convolution operation ($W - 1$ zeros are padded in each direction). Introducing a stride $s > 1$ in the convolution (2.1) corresponds to computing only every s-th element of the output feature map, which is often used to downsample the feature map.

As is typical in this field, in this thesis, we use the terms convolution and cross-correlation synonymously as a convolution corresponds to a cross-correlation with a flipped filter. CNNs further consist of nonlinear activation functions, that are needed to learn nonlinear relationships, e.g. widely-used are the Rectified Linear Unit (ReLU) functions, which are defined as

$$\mathrm{ReLU}(x) = \max(x, 0), \tag{2.2}$$

resulting in positive values only. Without nonlinear activation functions, a neural network would itself be a linear transformation and therefore be restricted to a linear regression model. Additionally, pooling layers typically succeed a convolutional layer in CNNs and reduce the resolution of the feature maps, which not only lowers the overall memory cost and performs dimensionality reduction but also counteracts overfitting. Examples of pooling functions are max pooling, min pooling or average pooling, that sample the maximum, minimum or the average of given values. Finally, a few fully connected layers, in which all neurons are connected to all neurons of the preceding layer, return the final class predictions. For an activation function σ, an input $x \in \mathbb{R}^M$, a weight matrix $w \in \mathbb{R}^{N \times M}$ and a bias $b \in \mathbb{R}^N$, the n^{th} output $y_n \in \mathbb{R}$ (for $n = 1, \ldots, N$) of a fully connected layer is given by

$$y_n = \sigma(w_{n,1}x_1 + \cdots + w_{n,m}x_M + b_n). \tag{2.3}$$

In this type of layer, each input of the input vector impacts each output of the output vector. Depthwise separable convolutions are simplified convolutional layers of CNNs, which can be evaluated at a lower computational cost and require fewer parameters. Architectures based on these layers were developed targeting mobile devices and are introduced in the next section.

2.1.1 Depthwise Separable Convolutions

In 2013, the concept of depthwise separable convolutions was initially developed at Google Brain by the intern Laurent Sifre to improve the classification performance of AlexNet [KSH12] while reducing the overall model size [Lau14; SM13; Cho17]. It decomposes the standard convolution, which performs channel- and spatial-wise computations in a single step, into two smaller parts, a depthwise convolution, followed by a pointwise convolution. Instead of convolving the filter over all input channels, during the depthwise convolution, a single kernel is only applied to one input channel. Then, the final pointwise convolution uses a 1×1 kernel to generate a linear combination of the output from the depthwise convolution. As described by [Guo+19; How+17], the depthwise separable

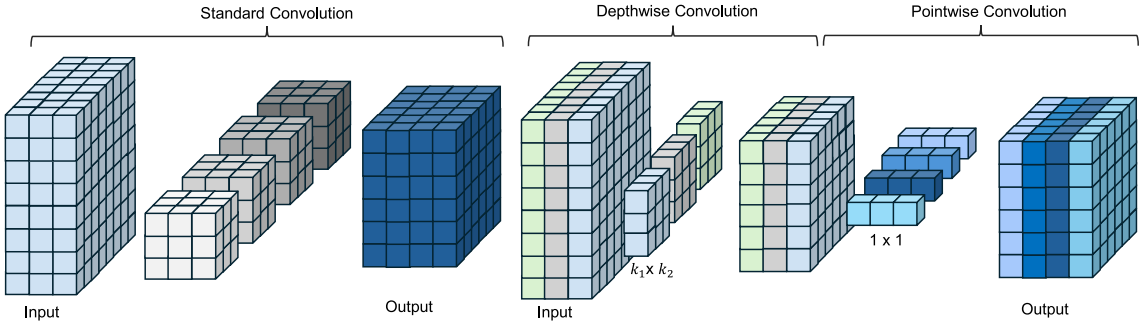
convolution is now composed of the depthwise convolution with the kernel $\widehat{K}$ of size $W \times W \times M$

$$\widehat{O}_{k,l,m} = \sum_{i,j} \widehat{K}_{i,j,m} \cdot F_{k+i-1,l+j-1,m}, \tag{2.4}$$

and the 1×1 pointwise convolution with the kernel $\widetilde{K}$ of size $M \times N$

$$O_{k,l,n} = \sum_{m} \widetilde{K}_{m,n} \cdot \widehat{O}_{k-1,l-1,m}. \tag{2.5}$$

The difference between a standard and a depthwise separable convolution is further visualized in Figure 2.1a and Figure 2.1b, respectively.



(A) Standard 2D Convolution. (B) Depthwise Separable Convolution.

FIGURE 2.1: Difference between a standard 2D convolution and a depthwise separable convolution.

This type of convolution requires significantly fewer parameters and thus, reduces the computational overhead. To elaborate on this, the required number of parameters and operations are computed for both convolution types in the following [How+17].

The number of parameters. A filter of a standard convolution consists of

$$P = W^2 \cdot M \cdot N$$

parameters. In contrast, the depthwise separable convolution requires two filters, each consisting of a separate set of parameters, which together contain

$$P_{\text{sep}} = W^2 \cdot M + 1 \cdot 1 \cdot M \cdot N$$

parameters. The ratio of necessary parameters for a filter of a full convolution and the filters of a depthwise separable convolution is therefore given by

$$\frac{P_{\text{sep}}}{P} = \frac{W^2 \cdot M + M \cdot N}{W^2 \cdot M \cdot N} = \frac{1}{N} + \frac{1}{W^2},$$

which shows that depthwise separable convolutions reduce the amount of necessary parameters in a particularly significant way when W and N are both relatively large.

The number of operations. A standard convolution with $D_F \times D_F$ input features for each of the M input channels and $D_O \times D_O$ output features for each of the N output channels, and a $W \times W$ kernel can be computed directly with

$$C = W^2 \cdot M \cdot N \cdot D_O^2 \tag{2.6}$$

operations. For depthwise separable convolutions, the number of parameters can be computed by

$$C_{\text{sep}} = W^2 \cdot M \cdot D_F^2 + 1 \cdot 1 \cdot M \cdot N \cdot D_O^2. \tag{2.7}$$

Overall, the ratio of necessary computations is therefore given by

$$\frac{C_{\text{sep}}}{C} = \frac{W^2 \cdot M \cdot D_F^2 + M \cdot N \cdot D_O^2}{W^2 \cdot M \cdot N \cdot D_O^2} = \frac{D_F^2}{D_O^2}\frac{1}{N^2} + \frac{1}{W^2}.$$

In the typical case that D_F and D_O agree, this ratio further simplifies to

$$\frac{C_{\text{sep}}}{C} = \frac{1}{N^2} + \frac{1}{W^2},$$

which is the same ratio as observed from the number of required parameters.

2.1.2 Batch Normalization

Batch normalization was first introduced in 2015 as an additional regularizer and accelerator technique during the training of deep neural networks by reducing internal covariate shift [IS15]. The internal covariate shift describes the variation of the distribution of the activations originating during training through constantly changing parameters. This demands that the layers readjust to the new values, which specifically for large, deep neural networks, is propagated and can accumulate during training, ultimately leading to a shift. Normalization of the means and variances of each layer's input applied to each batch mitigates these phenomena. Thus, for the variance σ^2 and the mean μ of each mini-batch, [IS15] proposed to normalize each dimension of a layer with d-dimensional input $x = (x^{(1)}, \ldots, x^{(d)})$ with

$$\hat{x}^{(k)} = \frac{x^{(k)} - \mu[x^{(k)}]}{\sqrt{\sigma^2[x^{(k)}]}}. \tag{2.8}$$

For each activation $x^{(k)}$, they further introduced the parameters $\beta^{(k)}$ and $\gamma^{(k)}$ to shift the normalized parameters with

$$y^{(k)} = \gamma^{(k)}\hat{x}^{(k)} + \beta^{(k)}. \tag{2.9}$$

Although this reasoning for the improvement by batch normalization is under discussion [San+18b; Bjo+18], this technique has since been incorporated in many deep neural networks and improves neural network training. A prominent example is presented in the following.

2.1.3 MobileNet Architectures

Following the development of depthwise separable convolutions, in 2017, [How+17] first presented a new model type, named MobileNet, which incorporates the aforementioned depthwise separable convolutions to form lightweight neural networks. This image classification model was introduced along with two hyperparameters, width and resolution multiplier, to vary the size of the MobileNets by the user and to thereby set individual trade-offs between accuracy and required resources. The MobileNet architectures use 3×3 depthwise separable convolutions, leading to a reduction up to 9 times compared to standard convolutions by achieving competing accuracies on ImageNet compared to other popular models [How+17]. Each convolutional layer succeeds a batch normalization layer and a ReLU. The average pooling is needed for a reduction of the spatial dimension to 1, preceding the final fully connected layer. Based on this, the MobileNetV2 [San+18a] evolved, which is characterized by building blocks including residual and linear bottleneck structures. Within these blocks, the depthwise separable convolutions begin with a pointwise convolution, followed by a depthwise convolution and then another pointwise convolution. Thus, in contrast to the standard depthwise separable convolution presented before, this architecture first increases the dimension of the feature map, before the channel-wise computation follows. In those building blocks, the input and output layer are further combined with a residual connection. This was further refined, resulting in the MobileNetV3 [How+19], which was introduced in 2019 and generated by employing various architecture search algorithms to optimize the MobileNetV2. In addition to the MobileNetV2 blocks, it consists of Squeeze-and-Excitation blocks, which compute channel-wise recalibration dynamically, enhance the representational functionality and further lead to performance improvements [HSS18]. To contextualize the relevance of efficient model architectures such as the MobileNets, Figure 2.2 based on [CPC16; BLB22] depicts the accuracy achieved by several deep learning models on ImageNet in comparison [CPC16].

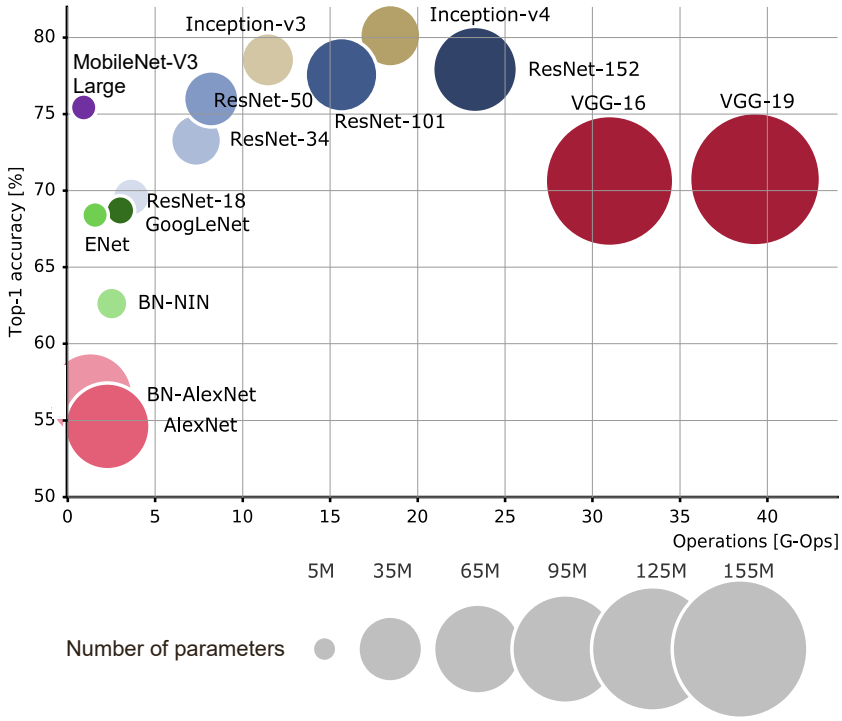


FIGURE 2.2: Top-1 accuracy of several CNN models on ImageNet over the years with the circle size corresponding to the model complexity based on [BLB22; CPC16].

The model complexity is thereby visualized by varying dot sizes which correlate with the number of parameters. The number of operations per second is additionally indicated. Handcrafted models such as the VGG-16 and VGG-19 often obtain a high accuracy, but become large quickly [Ben+21]. In contrast, efficient handcrafted models are designed to reduce the complexity while maintaining a given accuracy, a prominent example is given by the model Inception-v3. The MobileNetV3 was generated by hardware-aware Neural Architecture Search (NAS) (HW-NAS), which aims to optimize the accuracy and complexity jointly given a target hardware architecture. This resulted in a comparably very low model complexity and a competitive Top-1 accuracy on ImageNet.

In general, the MobileNet family was developed to perform classification locally on-device and thus, omit the necessity of transmitting data to an external server for final evaluation. This new model architecture allowed creating smaller and computationally faster CNNs while performing accurate image classification. Within this work, a variety of other neural network architectures for image classification tasks were tested and evaluated. The MobileNetV3-Small yielded the best results considering a trade-off of accuracy and model complexity. Hence, experiments conducted within this thesis build upon these findings and use the MobileNetV3 as the standard neural network architecture for image classification. With a state-of-the-art model comprising only a few million parameters, it consists of a suitable starting point for generating models targeting embedded low-power applications. Going forward, the size can be further reduced by various optimizing strategies.

2.2 Autoencoder

An autoencoder [RHW85; LF87; BK88] is a subclass of neural networks, that consist of an encoder-decoder framework. It typically comprises at least three layers, an encoding or input layer, a layer with the encoded data, which is also named the latent space, and the decoding or output layer. In so called undercomplete autoencoders, the encoder E compresses an input to remove potential noise and produces lower dimensional data, while the decoder computes a reconstruction D based on the compressed data [GBC16]. The reconstruction error of the autoencoder $D(E(x)) - x$, is penalized with a loss function

$$L(x, D(E(x))), \tag{2.10}$$

such as the Mean Squarred Error (MSE). During the training process of the autoencoder, such function L is minimized, e.g. by the gradient descent method. This training is unsupervised and can therefore be an effective way to encode information present in large unlabeled datasets, a situation that often appears in the context of medical applications. If the decoder simply consists of linear layer and the MSE is used, the learning of such undercomplete autoencoder is equivalent to learning to encompass the same subspace as a Principal Component Analysis (PCA) [GBC16]. When autoencoders comprise nonlinear encoder and decoder functions, they can potentially be trained to learn more complex data. Importantly, a very large capacity encompasses the risk that the model overfits and learns to simply copy the data without gaining relevant information on the underlying data distribution. Generally, the reconstruction error of the input and output data can be used to evaluate the autoencoder's capabilities. Thus, the main objective of an autoencoder is to perform dimensionality reduction by converting high dimensional data into lower dimensional data in a latent space.

2.3 Ensemble Learning

Ensemble learning uses a number of different learning algorithms, aiming to enhance the prediction performance compared to individual models [SK95; OM99; Rok10]. Multiple machine learning models are individually trained on the same task, and then the predictions combined to compensate individual weaknesses and obtain an improved result. For additional fine-tuning, the individual predictions can be weighted differently, when producing the final evaluation [Die00]. Originally, Bayesian averaging was employed as the first ensemble model. However, popular methods used in ensemble learning often include various neural networks and decision trees, such as bagging, random forest or boosted trees. Moreover, it was found that an increased diversity among different classifiers yields better results of the overall ensemble [KW03; Die00]. In this context, classifiers are considered diverse, if they produce different errors on unseen data points [Die00]. For example, with an ensemble of three classifiers $\{h_1, h_2, h_3\}$ and an unseen data point x, in the case of identical (not diverse) classifiers, when $h_1(x)$ produces a false classification, $h_2(x)$ and $h_3(x)$ are also incorrect. On the contrary, even if $h_1(x)$ is incorrect, the other classifiers might produce correct predictions, leading to a correct or more accurate prediction for the complete ensemble.

When using neural networks as base learners for an ensemble model, an objective is to introduce randomness to the learning algorithms to increase the diversity [Die00]. This can be realized by training multiple networks individually, since during backpropagation the initial weights of the models are determined randomly. Thus, even with the same training data, due to varying initialization, the models might lead to different results. A drawback of ensemble models is that the overall size might be larger than if single models are used. Furthermore, evaluating multiple individual models within the ensemble requires an increased evaluation time compared to simply using the individual models. However, in some contexts the enhanced classification performance outweighs the issue of needing additional computations. Finally, a key advantage of ensemble models is that supervised classifiers can be combined with unsupervised techniques, which can be used to address the class imbalances that are typical for medical data sets.

2.4 Random Forest Models

Random forest models [Bre01; Ho95] can be used for classification and regression tasks and were originally introduced as an extended form of the bagging [Bre96] algorithm. While in previous published methods, the nodes in the trees were split based on the best split of all variables, a random forest model chooses the best split among only a subset of prediction trees, which are randomly selected at each node. Thus, bagging can be considered as a special subclass of random forests, where the size of the subset is equal to the number of trees. In the following, the algorithm is outlined in more detail [Bre01; Bre02].

Starting at the root node, which contains all data, new trees are constructed based on a subset of the data and grown as large as possible. Given $f \in \mathbb{N}$ features and a parameter $k \in \mathbb{N}$, with $f < k$, at each node, k features are randomly selected targeting the best split, while searched through with all data available. A forest consisting of n trees, returns a classification prediction for an incoming sample x, based on the most frequent prediction of the uncorrelated subtrees for a classification problem. [Bre01] demonstrated that a strong correlation of various trees correlates with a worse performance. Thus, it is aimed to construct preferably uncorrelated trees. The main drawback of this method is an increased computation time the more trees are involved and the more depth the trees have. Compared

to deep neural networks, however, random forests require comparably little computational resources for training and inference. While random forest models can be chosen as an individual classifier within an ensemble model, they can be classified themselves as an ensemble learning method, since all individual subtrees are combined to a 'forest of trees', equivalent to an ensemble.

2.5 Support Vector Machines

Support Vector Machines (SVM) [CV95] belong to the supervised machine learning classifiers and are applied to solve classification as well as regression tasks by mapping input vectors nonlinearly to a feature space of high dimensionality. The mapping creates a hyperplane within the transformed feature space for non-separable training data. This hyperplane functions as a decision boundary between different classes. Depending on the feature dimensionality, the hyperplane ranges from a linear function for two-dimensional data or a plane for n-dimensional input data. A hyperplane is considered optimal, when the margin between the plane and data points is maximal. When the data is separable, a hard margin is generated. However, in the case of noisy data, precise separation is difficult, such that a soft margin decision boundary is applied. Let x_i be some data points in some vector space (say $\mathbb{R}^d$ for some d) with the labels $y_i \in \{-1, 1\}$ for $i = 1, \dots n$. As described in [SS02, Section 1.4], the parameters of a linear SVM are determined by the constrained minimization problem

$$\begin{aligned} &\min_{w,b,\zeta} \left[\tfrac{1}{2}\|w\|^2 + C \sum_{i=1}^{n} \zeta_i \right], \\ &\text{subject to} \quad y_i(\langle w, x_i \rangle + b) \geq 1 - \zeta_i \\ &\text{and} \qquad\qquad\quad \zeta_i \geq 0 \qquad \text{for all } i = 1, \dots, n, \end{aligned}$$

with the normal vector to the hyperplane w, the slack variable ζ_i and the positive hyperparameter C. The brackets $\langle \cdot, \cdot \rangle$ denote the standard scalar product. The present constrained minimization problem is formulated for linear SVM. For nonlinearly separable data the above formulation is used where the standard scalar product $\langle \cdot, \cdot \rangle$ is replaced by a more general kernel function $K(\cdot, \cdot)$, which is typically still symmetric and positive semi-definite, but no longer linear in each argument. This is equivalent to transforming the lower dimensional input vectors into feature vectors in a higher dimensional space and then applying the linear SVM, which is known as the kernel trick [BGV92]. Popular kernel functions are for example the Gaussian or radial basis function $K(x, y) = \exp\left(\frac{-\|x-y|_2}{2\sigma^2}\right)$ or polynomial kernel functions $K(x, y) = \left(x^T y\right)^d$ for appropriate choices of the variance $\sigma \in \mathbb{R}$ and the polynomial degree $d \in \mathbb{N}$, respectively. The inference of a SVM classifier corresponds to computing

$$f(x) = \operatorname{sgn}\left(K(w, x) + b\right).$$

In contrast to deep neural networks, SVMs do not rely on a large amount of labeled training data for proficient performance and might be more easily deployed on-chip.

2.6 Hardware Architectures

2.6.1 Single Instruction Multiple Data Stream Architectures

Flynn's taxonomy [Fly72] describes four different computer architectures: Single Instruction Multiple Data Streams (SIMD), Single Instruction Stream Single Data Stream (SISD), Multiple Instruction Streams Multiple Data Streams (MIMD) and Multiple Instruction Streams Single Data Stream (MISD). Prominent SIMD examples are vector architectures, multimedia extensions and Graphics Processing Units (GPU), which are essential for training neural networks and thus, are outlined in more detail in the following. They can be used for fast and parallel processing of multiple different data streams using only a single instruction [HP11]. Besides parallelism, another advantage of SIMD is given by a reduced instruction bandwidth and space since only a single copy of the instruction code is needed [PH16]. While each processing element possesses its own data memory, for multiple processing elements only a single instruction memory exists. Among others, this architecture is well-suited for processing image data, since the required operations for multiple images and pixels is repetitive. With SIMD, such instructions can be executed in parallel by using various functional units at once.

2.6.2 Systolic Array of MAC Units

A MAC-array with n rows and m columns consists of $n \times m$ MAC units and can execute multiple dot products in parallel [Sze+20]. They are typically employed in hardware accelerators and can be understood as a combination of SIMD and MISD. A MAC unit is a hardware unit that multiplies two numbers and adds their product to an accumulator. MAC-arrays enable parallel and fast matrix multiplications directly on an accelerator as required for applications exhibiting real-time requirements. They allow fast execution of neural networks by computing the majority of operations efficiently as vector-matrix multiplications.

2.6.3 Key Hardware Metrics

When hardware simulation of machine learning models is performed, different hardware metrics are used to assess these results. These are discussed in the following.

Number of Operations: MAC units can perform the most essential operations required for the convolutional layer of deep neural networks: a multiplication of two numbers followed by an addition, which accumulates the partial sum that is then stored in a local register. Specifically in the ML context, such MAC operation is considered as a single operation instead of two (multiplication and addition). For each layer, many computations need to be performed, which can be accelerated by exploiting parallelism of the MAC operations in order to increase the throughput. Reducing the word width of the data lowers the hardware costs of MAC units, such as the area and power. Overall, the number of MAC operations can be consulted as one additional metric to assess the hardware suitability of ML models.

Throughput: Throughput of an architecture means the amount of data that can be processed in a specific amount of time, more specifically the number of operations executed per second. Thus, the throughput describes the rate at which workloads are processed over a defined time period. It is affected by the hardware's parallelization, clock frequency and utilization capabilities. The utilization is limited by the target hardware itself and is given by the in theory possible execution time over the actual required execution time. [PH16; JV23]. For the VCE and the seizure detection using neuroimplants, the execution of

the main operation is conducted by MAC units. The peak throughput of a MAC unit can be enhanced by an increased clock frequency, a shortening of the critical path or an increased number of operations per cycle [Sze+20]. Furthermore, the throughput could be enhanced by an increased total number of MAC units. However, employing a larger number of MAC units, requires an increased area on hardware and thus enlarges the overall area cost. Concurrently, an enhanced throughput due to an increased number of parallel processing elements executed with a high frequency, also leads to an increased power consumption.

Latency: Latency defines the time delay between receiving data and actually returning a result, for example, the delay between capturing an image in the GI tract and the information of the final classification result. Especially for real-time applications, such as the VCE or seizure detection, a latency below given time constraints is essential. This is even more important for the seizure detection task, since neurostimulation needs to be applied directly after a beginning seizure is detected. Furthermore, an increased throughput correlates with a decreased latency. [PH16; JV23].

Power Consumption: The term power consumption describes how much electrical energy is consumed over a specific unit of time and is measured in Watts W or joules per second J/s. This is predominantly increased by an enhanced data movement, especially if the distance between the stored data and a MAC unit is large, since data movement over a larger distance demands more energy than over a smaller distance. Reducing data movement and lowering the number of memory accesses is essential as it generally results in a reduced power consumption.

Energy Demand/Energy Consumption: While the power consumption only determines how much power is needed for a single moment, electric energy consumption can be described as the integral of the power consumption over a defined time frame and is measured in Joules. Thus, it describes the total amount of electric energy E used across a time t with

$$E = \int P(t)\, dt. \tag{2.11}$$

Optimizing the processing order of data and data reuse, in which the same data is used for multiple consecutive operations, can help to reduce the power and energy consumption. For example, data can be loaded temporarily from a costly large and distant memory, e.g. DRAM, into a small, local on-chip memory and discarded only after all operations at that time have been conducted [Sze+20]. Reducing the cost of moving data is another option to lower the overall energy demand and can be for example realized by reducing the data word width. Thus, quantizing the networks with a lower word width is another important objective when designing and implementing adequate machine learning models.

Energy Efficiency: The term energy efficiency describes the number of operations executed while consuming a given unit of energy [PH16; JV23]. It evaluates how much energy is needed per operation (assessed in J/ops) or for a single inference. Depending on the task complexity and the total energy demand, a high energy efficiency can be important when designing hardware-aware model for sensor edge devices, such that an interplay of various factors needs to be considered. Therefore, when designing machine learning models for low-power edge devices, this can be an important factor.

Data Width: Many low-power hardware architectures generally only employ fixed-point computations, which makes its usage indispensable when designing machine learning

models targeting embedded devices. Computation with this representation is additionally significantly faster and more energy-efficient than using floating-point representation [PH16]. With this computation, the radix point (binary point) is fixed at a specific position within the data word and if addition or subtraction operations are used, the radix point is not changed. However, in multiplication and division cases, it possibly needs readjustment. Nevertheless, employing a 32- or 64-bit floating-point representation is still more common in the ML community [Doc] during neural network training rather than using a 32-, 16- or 8-bit fixed-point representation. Notably, a decreased precision due to fewer bits per operation directly results in a reduced memory bandwidth.

Memory Consumption: The memory consumption describes the amount of memory space, which is needed for a program. It can be considered as the data width · number of elements (e.g. weights or features of a neural network, number of instructions). Since the memory space of small edge devices is generally limited, this also needs to be reflected when choosing adequate hardware and ML models.

Area: The chip area is typically reported in mm^2 and contributes significantly to the monetary cost of the hardware. It is predominantly limited by the targeted application and the technology. The employed model size also directly impacts the required on-chip storage size.

Ultimately, the usage of neural networks must be supported by the hardware, which has to be designed with an efficient mapping and adequate dataflows making efficient usage of MAC units and optimized memory accesses [Sze+20]. Nevertheless, when conceptualizing AI-based decision pipelines targeting low-power embedded devices, choosing adequate and hardware-friendly, lightweight, neural networks from the start is important as it positively impacts key hardware metrics such as the overall energy demand and should be considered as early as possible. As a first step, the number of MAC operations and the total model size on their own might provide an initial reference to assess the hardware suitability of a model. However, to evaluate the energy demand, more factors, such as power consumption, clock frequency and inference time, need to be considered and a comprehensive hardware-software co-design is beneficial.
This work aims to present the number of MAC operations along with the number of weights for purely software-based experiments. Furthermore, some selected and promising models should be trained with quantization-aware training to reduce the word width and the computational demand while still maintaining a high performance. Finally, hardware simulations should be conducted if possible for both medical applications.

2.7 Time Series Analysis

Time series analysis methods are used to gather more information about a problem or task from time-series data. In the following, HMMs and their underlying Markov chains are introduced, followed by an explanation of Viterbi Decoding.

2.7.1 Markov Chains

Markov Chains were invented by Andrey Markov in 1906 [GS12]. They are a simple time series models, where the distribution of a future state depends only on the current state. A Markov chain is a sequence of random variables $X_0, X_1, X_2, \ldots$ with values in a discrete set $S = \{s_1, \ldots, s_n\}$, the state space. The distribution of Markov chains fulfill two key properties: firstly, for all $t \in \mathbb{N}$ and all paths $x_0, \ldots, x_{t+1} \in S$, we require

$$\mathbb{P}(X_{t+1} = x_{t+1}|X_t = x_{t,\dots}, X_0 = x_0) = \mathbb{P}(X_{t+1} = x_{t+1}|X_t = x_t). \tag{2.12}$$

The distribution of the future state X_{t+1} therefore only depends on the current state $X_t = x_t$, but not on any previous state $(X_{t-1} = x_{t-1}, \dots, X_0 = x_0)$.

Secondly, these transition probabilities are further assumed to be constant in the sense that they do not depend on t, namely we have

$$\mathbb{P}(X_{t+1} = s_j|X_t = s_i) = a_{ij} \quad \text{for all } t \geq 0, \tag{2.13}$$

with some transition probabilities a_{ij}, for $i, j = 1, \dots, n$. These transition probabilities are collected in the matrix $A = (a_{ij})_{i,j=1,\dots,n}$, fulfill $\sum_j a_{ij} = 1$ and are positive $a_{ij} \geq 0$. The model is completed by a given distribution α of the initial state, namely we set

$$\mathbb{P}_{X_0} = \alpha.$$

This determines the total probability distributions of all X_t, e.g. for X_1, we can directly compute

$$\mathbb{P}(X_1 = s_j) = \sum_i \mathbb{P}(X_1 = s_j|X_0 = s_i)\mathbb{P}(X_1 = s_i) = \sum_i \alpha_i a_{ij} = (\alpha^T A)_j, \tag{2.14}$$

for $j = 1, \dots, L$. Here, $(\alpha^T A)_j$ denotes the j-th entry of the row vector that is obtained by the vector-matrix product $\alpha^T A$.

2.7.2 Hidden Markov Model

A Hidden Markov Model (HMM) is an extension of the aforementioned Markov chain, in which the states are hidden and cannot be directly measured, but have to be inferred from emissions [BP66; Rab89; Edd96]. HMMs are beneficial when analyzing time-dependent data and compared to many machine learning models, such as deep neural networks, present models of low complexity and are widely used in many fields for various applications. For example, they are a popular tool for weather forecasting [Ail+15] and speech recognition [Ren+94; Gal98]. In bioinformatics, HMMs are employed for the alignment of Deoxyribonucleic Acid (DNA) sequences or protein sequences [Chu89; Söd05; Rem+12] and for the detection of CpG islands (cytosine and guanine rich regions in the genome) [Wu+10]. While the (hidden) states themselves are not visible in a HMM, their emissions can be observed and based on the emissions, estimators for the hidden states can be derived. The emission probability $\beta_i(b)$ for each $s_i \in S$ and $b \in O$ with an alphabet O of observations is defined as

$$\beta_i(b) = \mathbb{P}(O_t = b|X_t = s_i). \tag{2.15}$$

Here, O_t and X_t are the random variables corresponding to the emission and hidden state at the moment t respectively. We note that the emission probabilities are independent of the moment t, the above equation is therefore assumed to hold for all $t \in \mathbb{N}$, $s_i \in S$. Figure 2.3 provides an example of a HMM, with n hidden states S, 3 different observations O, the transition probabilities α and emission probabilities β. By observing a sequence of

emissions, likelihoods of paths of the hidden states can be computed and the solution of a dynamic programming problem gives the maximum likelihood estimator, which is known as Viterbi decoding.

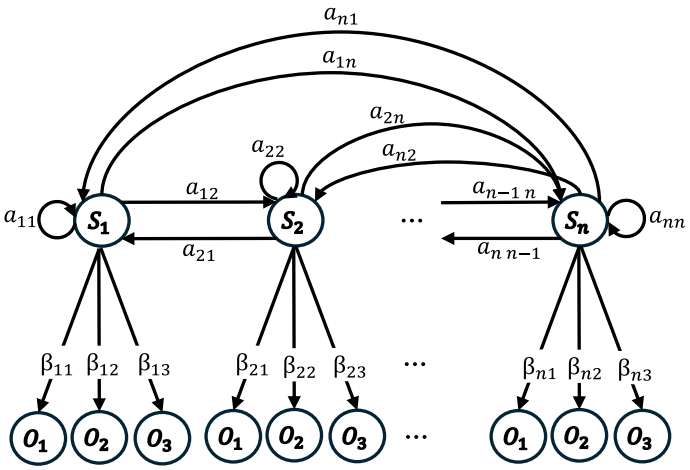


FIGURE 2.3: Example of a Hidden Markov Model.

2.7.3 Viterbi Algorithm

The Viterbi algorithm was originally developed by Andrew Viterbi in 1967 [Vit67] to decode noisy convolutional codes in the communication field. Since then, its usage spread across multiple subjects. Especially in bioinformatics, the Viterbi algorithm is applied to solve DNA sequence alignments or for protein structure prediction by estimating the most likely protein conformation based on an amino acid sequence [Dur+98]. Based on dynamic programming techniques, the Viterbi algorithm computes the most likely state sequence of a HMM, given a sequence of emissions [For73]. More specifically, given a HMM and a sequence of observations $o_1, \ldots, o_t$, the objective is to find the most likely sequence of hidden states $x_1, \ldots, x_t$ with

$$\arg\max_{x_1,\ldots,x_t} P(X_1 = x_1, \ldots, X_t = x_t, O_1 = o_1, \ldots, O_t = o_t), \tag{2.16}$$

to solve this decoding task. The Viterbi algorithm consists of three main steps: initialization, a recursion phase that builds up a likelihood matrix v and a sequence of backtracking pointers p, and finally a backtracking phase that recovers the estimated path of the hidden states. First, the Viterbi algorithm is initialized by setting, for each state, the first cell of the so-called trellis to $v_1(j) = \alpha_j b_j(o_1)$ and the initial backtracking pointers to $p_1(j) = 0$ for all $1 \leq j \leq n$.

Next, the likelihood matrix v_t is recursively build up and computes, during each time step, the likelihood of the most likely path up to the observation to this time step t under the assumption that the Markov chain is in the state j at time step t, namely the expression

$$v_t(j) = \max_{x_1,\ldots,x_t} P(X_1 = x_1, \ldots, X_{t-1} = x_{t-1}, O_1 = o_1, \ldots, O_t = o_t, X_t = j). \tag{2.17}$$

Since the Markov chain is oblivious, in the sense that its distribution of time step t only depends on the state of the chain at time step $t-1$, this likelihood simplifies to the expression

$$v_t(j) = \max_{i=1}^{n} v_{t-1}(i) a_{ij} b_j(O_t), \tag{2.18}$$

for $1 \leq j \leq n, 1 < t \leq T$, until some final time step T. The cell indicating the backtracking pointer stores the information of the actual path that realizes these likelihoods and can be recursively computed alongside v_i via

$$p_t(j) = \arg \max_{i=1}^{n} v_{t-1}(i) a_{ij} b_j(O_t); \quad 1 \leq j \leq n, \ 1 < t \leq T. \tag{2.19}$$

For T observations with a state space consisting of n different states, this algorithm requires the storage of $\mathcal{O}(T \cdot n)$ parameters and has a time complexity of $\mathcal{O}(T \cdot n^2)$. Thus, this approach is well-suited to determine patterns within hidden sequences of time-series data based on a series of observations.

2.8 Electroencephalography

The Electroencephalography (EEG) was first introduced by Hans Berger [Ber29] in 1929. It is a non-invasive medical diagnostic method, that measures the brain activity by detecting voltage differences of the pyramidal neurons with electrodes. The electrodes are uniformly attached along the scalp, typically according to the established International $10-20$ system [Cob+58], referring to the relative distance of adjacent electrodes in relation to a defined reference node as illustrated in Figure 2.4.

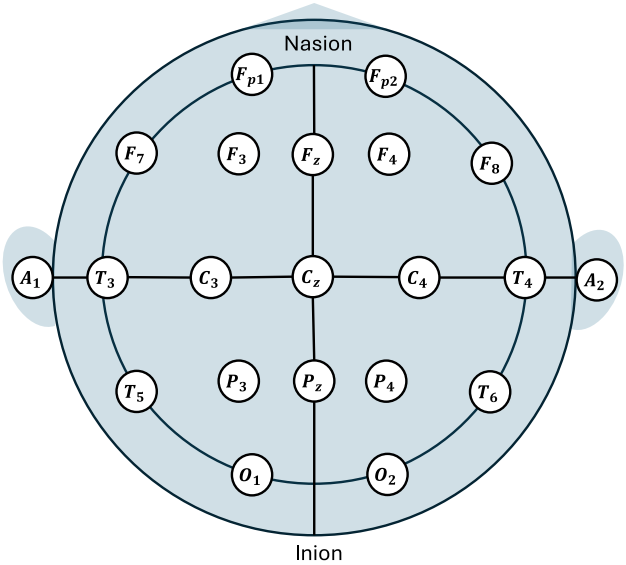


FIGURE 2.4: Electrode distribution according to the International $10-20$ system for an EEG measurement.

With such an arrangement, the voltage between two electrodes can be measured, which are generally between $20-100\mu V$ [Aur+04], if measured on the scalp. There are some distinct brain waves, which occur in the background of an EEG and can be classified based on their frequency [RI17]. However, it should be noted that the boundaries for classification

of those band waves vary depending on the definition. In the following, they are outlined based on [SC13; BBA20].
The delta brain waves are the slowest with $1 - 4$ Hz with high amplitudes and are usually observed during deep sleep in adults. Theta brain waves are characterized by an amplitude $\leq 100\,\mu V$ and a frequency of $4 - 8$ Hz, typically present in children and sleepy individuals. Alpha waves ($8 - 12$ Hz) have an amplitude of $\leq 50\,\mu V$ and occur in a relaxed state or with closed eyes while focused awareness and open eyes lower its amplitude. If a person is focused or actively moving, beta waves with a frequency of $12 - 25$ Hz and an amplitude $\leq 30\,\mu V$ can be usually observed. Finally, wave bands with a frequency ≥ 30 Hz and typically the lowest amplitude are labeled as gamma waves, most predominantly present in a variety of cognitive states, including active memorizing.

It has been reported that differences in the amplitude of EEG data can already be detected seconds before the seizure actually begins [JMM21]. If this preictal (pre=before, ictal=seizure) phase was detected in real-time, seizures can be classified and optimally treated before harmful consequences can occur. Figure 2.5 displays the EEG seizure (ictal) pattern in comparison to a preictal, interictal and postictal signal.
The specific EEG channels involved in an increased electric activity during a seizure are very individual and can vary across different patients [Sho09]. Figure 2.6a and Figure 2.6b provide an example of how different a seizure recording for two individuals based on a scalp EEG might be.

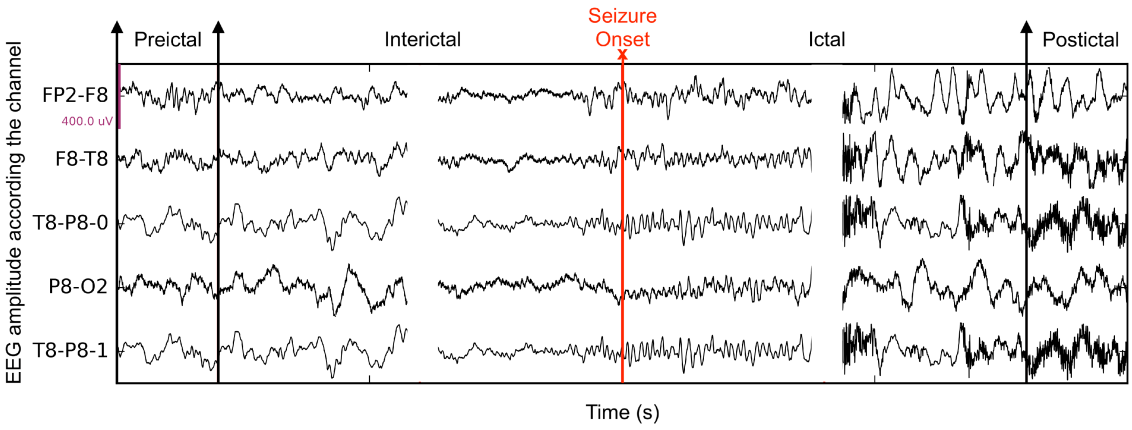
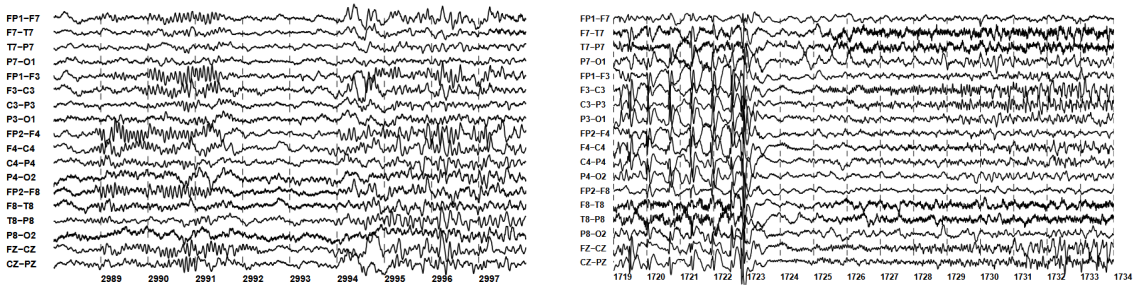


FIGURE 2.5: EEG amplitude in µV for 5 different channels and seizure phases over time, the seizure onset is marked in red, from [JMM21].



(A) Scalp EEG of a seizure from patient A.

(B) Scalp EEG of a seizure from patient B.

FIGURE 2.6: Example Scalp EEGs for two patients, from [Sho09].

For patient A, after 2994 s a seizure occurs, which is most strikingly visible in the channels FP2-F4 and T8-P8. In contrast to that, for patient B, the seizure starts after 1723 s and is most prominent in channel F3-C3. Importantly, one has to differ between an actual seizure

and normal sleep EEG features. For patient A between $2989 - 2992$ s, a spindle can be observed, which is a normal sleep feature of high oscillatory activity in mammals. While seizures vary across individuals, for the same person, the recordings with seizures show recurrent patterns. Thus, for this application retraining models with patient individual data might be beneficial for an enhanced prediction performance.

Challenges arising from EEG data
Several challenges accompany EEG data acquirement and analysis. EEG data that is acquired subcutaneously or through scalp recordings exhibit, in comparison to more invasive measurements, a low signal-to-noise ratio. Potential artifacts are mainly generated by physiological noise due to body movements and muscle contractions, the heart beet or eye movements, even a single muscle change can result in a change of the brain activity in the cortex. Additionally, technical artifacts, such as power line interference, ventilation or faulty electrodes influence the received recordings [Ras+20; Ein+23]. Theoretically, body-generated artifacts can be reduced by invasive EEG recordings. However, although there are possibilities of invasively acquiring EEG data, scalp recordings are much more common and overly preferred. Compared to invasive acquisition, scalp recordings are easier to obtain and more importantly, less harmful, as they do not require surgery. Furthermore, if the main objective is the design of AI-based models for neuroimplants, which are implanted on the scalp, it might be beneficial to train the model to filter out noise, which is inherently present on the scalp as well, although less than on the skin. Thus, the removal of artifacts is a crucial step in preprocessing of EEG data. Moreover, techniques to analyze EEG data are not standardized, leading to varying labeling of different experts for some instances. In addition, as healthy and abnormal EEG recordings can differ between individuals, it is difficult to assess what constitutes normal activity.

2.8.1 EEG Dataset - Children's Hospital Boston - Massachusetts Institute of Technology (CHB-MIT)

The training of machine learning models requires datasets containing real-world data from patients for both applications. In general, successful seizure detection based on EEG data relies on actual EEG data from patients suffering from seizures, including healthy recordings along with seizures for each patient. These datasets often contain large amounts of healthy data and in comparison, only a small fraction of seizure recordings. There are multiple EEG datasets published, for example the Epilepsiae database [Ihl+12] as one of the largest datasets with recordings from more than 250 patients, which is only available by payment. Furthermore, there is a small, free of charge and publicly available EEG dataset from Bonn [And+01] and additionally, a larger, public EEG database: the Temple University Hospital (TUH) EEG dataset [OP16]. Lastly, the publicly available CHB-MIT dataset [Sho09; Gut10] is one of the most commonly used databases in the literature [Isl+24; Sho+21; Tru+18; UHA+18] and thus qualifies well as a comprehensive database for model evaluations in this work. It contains EEG recordings from 24 pediatric patients with epileptic seizures is described in the following.

For each patient, continuous EEG data was sampled at 256 Hz, meanwhile no anti-epileptic drug therapy was applied. In total, 664 European Data Format (EDF) files consisting of EEG data with 198 seizures and at most one seizure per EDF were acquired. For this dataset no official split was provided, leading to a variety of different splits employed by different authors, which complicates the comparability between different works. The applied preprocessing in this work is therefore described in detail to enable the reproduction of this concrete setting. Figure 2.7 illustrates how the training, retraining and test sets are generated from the original dataset and which proportions are used.

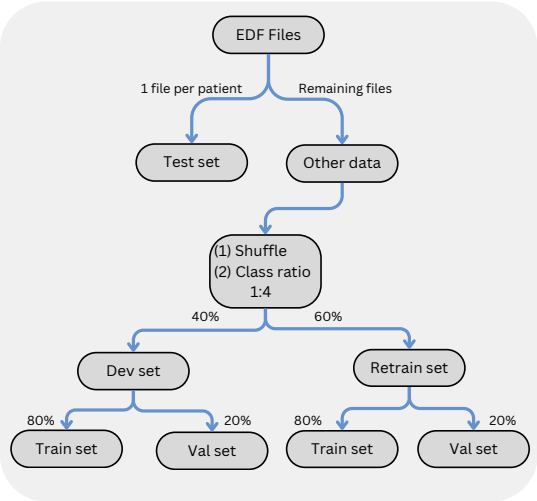


FIGURE 2.7: Allocation of the CHB-MIT dataset into development (dev), retraining (retrain) and test set.

First, for each patient, randomly a single file was excluded for the test set, while the remaining files were split into a development (dev) and a retraining set with a ratio of 40:60. Both sets are further split into a training (train) and validation (val) set (80:20). The retraining set was generated, since the process of retraining has proven to improve the classification performance for individual patients. The EEG data in this dataset includes data points that were generated from a varying number and type of EEG channels. Thus, for a uniform representation, a pre-selection of common channels needs to be performed prior to neural network training as described in Section 3.1.1.

2.9 Video Capsule Endoscopy - Datasets

To train image classifiers for VCE applications, it is not advisable to directly rely on datasets generated from classical colonoscopies or esophagogastroscopies, as they do not contain images from the small intestine. However, images retrieved from the small intestine are essential for VCE applications as characteristics, such as intestinal villi of the mucosa required for the absorption of nutrients, are not present in the rest of the GI tract. Moreover, data acquired from VCE studies is characterized by a lower image resolution and decreased framerate compared to data obtained from standard techniques [Sme+21]. Using images with a higher resolution compared to data originating from miniature camera sensors, as present in VCEs, impairs the validity of trained models and is avoided in this work. Additionally, in contrast to traditional colonoscopies, during the VCE procedure, the visibility is not improved by air inflation. Furthermore, controlling the capsule's movement is not possible in contrast to manual examinations. For those reasons, this work only relies on data generated from VCE for the training of image classifiers. Subsequently, the three largest VCE datasets were used to train image classifiers: the Kvasir-Capsule dataset [Sme+21], the Rhode Island Gastroenterology dataset [Cha+22], and the Galar dataset [Le +25].

2.9.1 VCE - Kvasir-Capsule Dataset

The Kvasir-Capsule dataset [Sme+21] was published in 2021 as the first publicly available VCE dataset and consists of 4,741,504 frames that were generated from 117 VCEs, with the Olympus Endocapsule 10 System. However, only 47,238 images of 43 patients in total were annotated in the process, resulting in 4,694,266 unlabeled images from the remaining 74 studies. The annotated frames were labeled for 14 different classes, including three anatomical landmarks: pylorus (a sphincter, controlling the emptying of the stomach into the duodenum of the small intestine), ampulla of vater (duct between the duodenum and the gall duct) and ileocecal valve (transition between small bowel and colon, limiting reflux of colonic contents). Luminal findings, such as the foreign body, reduced mucosal view (e.g. introduced by bubbles or bowel content), and normal clean mucosa account for the next three classes. Table 2.1 depicts the total number of annotated images for each of the 6 classes.

Table 2.1: Statistics of the Kvasir-Capsule dataset from [Sme+21] including the number of frames for the anatomical landmarks and luminal findings.

Pylorus	**Ampulla of Vater**	**Ileocecal Valve**	**Normal Mucosa**	**Reduced View**	**Foreign Body**
1,529	10	4,189	34,338	2,906	776

Finally, the dataset consists of annotations for the pathologies angiectasia (dilated vessels leading to chronic bleeding and potentially anaemia), erosion (lesion of the mucosa), erythema (redness of mucosa), fresh blood, hematin blood, lymphangiectasia (dilated lymphoid vessels), polyp (protrusion in the mucosa, potentially precancerous) and ulcer (large erosions). The number of labeled images for each pathology is presented in Table 2.2.

Table 2.2: Statistics of the Kvasir-Capsule dataset from [Sme+21] including the pathologies.

Angiectasia	**Erosion**	**Erythema**	**Fresh blood**	**Hematin Blood**	**Lymphangiectasia**	**Polyp**	**Ulcer**
866	506	159	446	12	592	55	854

While the smallest class (hematin blood) merely consists of 12 frames, the largest class (healthy mucosa) includes 34,338 annotated images in total. The acquired data of the remaining classes additionally varies. This emphasizes a huge class imbalance, not only within all data, but additionally between normal and pathological samples, which should be paid attention to, when searching for well-performing classification models.

As the authors emphasize, when AI-based classification models for VCE images are designed, splitting the VCE datasets into training, validation and test set needs to be carefully conducted and with the awareness that no related frames of the same patient should appear across the different datasets [Sme+21]. Optimally, the splits should be generated on a patient-level, such that each patient study can only appear in one of the datasets (training, validation or test). Accordingly, for the Kvasir-Capsule dataset, the authors provide such splits for the division of training, validation and testing data without incorporating information of the same patient across the different subsets. Furthermore, they excluded the smallest classes, namely hematin blood, ampulla of vater and polyp. For best reproducibility, the provided official splits by the authors were adopted for all experiments in this work.

2.9.2 VCE - Rhode Island Dataset

One of the largest publicly available VCE datasets is the Rhode Island Gastroenterology Dataset, which was published by [Cha+22] in 2022. This dataset consists of 5,247,588 VCE images from 424 patients in total, captured with a PillCamTM SB3, each labeled as one of the anatomical organs of the GI tract: esophagus, stomach, small intestine or colon. As many medical datasets, this dataset is also characterized by a class imbalance. Table 2.3 shows an overview of the descriptive statistics of the Rhode Island (RI) dataset with the total number of images per organ.

TABLE 2.3: Statistics of the RI dataset with the total number of images per anatomical organ **before** downsampling as presented in [Cha+22].

	Esophagus	**Stomach**	**Small Intestine**	**Colon**
Minimum	1	1	1,737	1
Maximum	3,152	25,081	37,240	31,184
Mean	32	1,314	9,698	1,332
Total	13,715	557,049	4,111,865	564,959

For individual studies, there is a discrepancy between the minimal number of frames in the largest class, small intestine, with 1,737 frames, and 1 for the other three organs. The total number of images across all patient studies varies significantly between the different organs, with a majority of images being captured in the small intestine, which make up for more than 4 million frames in the dataset. In contrast to that, only about 13,000 images were gathered from the esophagus. This demonstrates the class imbalance, which is omnipresent in VCE images.

Thus, the authors performed downsampling on the largest classes within the training and validation set, to reduce class imbalances and limited the required computational resources for performing training of neural networks with this data (see Figure 2.4).

TABLE 2.4: Statistics of the RI dataset with the total number of images per anatomical organ **after** downsampling as presented in [Cha+22].

Dataset	**Esophagus**	**Stomach**	**Small Intestine**	**Colon**	**Total Images**
Train/Val raw	9,061	466,562	3,242,639	474,776	4,193,038
Train/Val downsampled	9,061	11,508	32,252	11,707	64,528
Testing	4,654	90,487	869,226	90,783	1,054,550

While all frames from the esophagus class were maintained in the development set, only 1/400 for the small intestine, and 1/100 for the colon and stomach were randomly sampled in each study. The test studies were kept untouched, to simulate preferably realistic testing conditions. This modified dataset provided by [Cha+22] was used in the experiments conducted in this thesis and the associated publications without further modifying the proportions.

2.9.3 VCE - Galar Dataset

Most recently, in 2025, the Galar dataset [Le +25] was published, consisting of 3,513,539 annotated frames based on 80 VCE studies. This is the first publicly available VCE dataset, which comprises functional, anatomical and pathological annotations, making it particularly relevant for the problem at hand. The VCE data was acquired either with the Olympus™52 Endocapsule 10 System, PillCam™ SB2, PillCam™ SB3, or the Colon2

Capsule Endoscopy Systems. Thus, this multi-system-based dataset presents a more diverse representation of VCE studies in terms of multi-labeling and different endoscopy systems than the preceding datasets, counteracting potential domain shifts. More specifically, this multi-label database comprises annotations for technical classes: good view, reduced view (of at least 50%), no view (reduced view of at least 95%), bubbles, dirt and anatomical classes, such as z-line (transition from esophageal to gastric epithelium), pylorus, ampulla of vater, ileocecal valve, mouth, esophagus, stomach, small intestine and colon. Table 2.5 shows the number of images available in the technical and the anatomical classes.

Table 2.5: Statistics of the Galar dataset with the total number of images per technical and anatomical classes [Le +25].

Good View	**Reduced View**	**No View**	**Bubbles**	**Dirt**	**Ampulla of Vater**	**Ileocecal Valve**
22,015	10,934	3,226	21,689	15,353	1,740	3,692
Z-line	**Pylorus**	**Mouth**	**Esophagus**	**Stomach**	**Small Intestine**	**Colon**
122	3,183	2,009	2,256	254,994	1,375,918	1,878,361

Furthermore, in addition to a normal class, the frames were annotated for the pathological classes ulcer, polyp, active bleeding, blood, erythema, erosion, angiectasia, Inflammatory Bowel Disease (IBD), foreign body, hematin, cancer and lymphangiectasia. The number of frames per class are listed in Table 2.6.

Table 2.6: Statistics of the Galar dataset with the total number of images per pathological class and for the normal class [Le +25].

Ulcer	**Polyp**	**Active Bleeding**	**Blood**	**Erythema**	**Erosion**	**Angiectasia**
11,428	18,415	5,325	391,715	6,228	39,105	16,803
IBD	**Foreign Body**	**Hematin**	**Cancer**	**Lymphangiectasia**	**Normal**	
4,929	94	31,773	11,681	17,660	2,958,383	

As indicated for the Rhode Island and the Kvasir-Capsule dataset, the Galar dataset is also characterized by a large class imbalance. Approximately 84.20% of images present normal mucosa within the GI tract, while only $\approx$ 15.8% accounts for pathologies. Furthermore, due to the natural anatomy of the GI tract, the anatomical classes mouth and esophagus are significantly smaller than the larger classes of small intestine and colon (e.g. the number of frames in the mouth classes accounts for only 1.07% of the classes of the colon class). Although the majority of images are labeled as normal, they typically originate from patients, who suffer from conditions that ultimately lead to the VCE procedure, but may not always present standard healthy mucosa. This can negatively influence the neural network training by introducing an inaccurate baseline of healthy samples. The Galar dataset is the VCE dataset that contains the highest number of annotated pathologies. To assess in which part of the small intestine most pathologies were found, the likelihood of the occurrence of an anomaly within the small intestine, (averaged over all 80 patient studies) is visualized in Figure 2.8.

Anomalies here encompass the pathologies active bleeding, angiectasia, erosion, ulcer, blood, erythema and polyp. If either one of these classes occurs in an image, this is considered an anomaly. For all patients, the position of anomalies was normalized over the total length of the small intestine by dividing by the total number of images found in the intestine for this patient. Finally, the likelihood was plotted for all patients. It can be observed that anomalies cluster increasingly towards the end of the small intestine. This emphasizes the importance of enabling the coverage of the whole small intestine with a VCE capsule

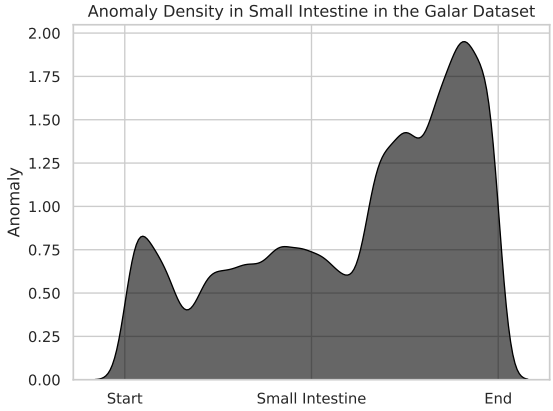


FIGURE 2.8: Anomaly likelihood over all 80 patient studies found in the Galar dataset (position of anomalies normalized over the total length of the small bowel).

before the capsule's energy is depleted. Therefore, designing hardware-aware ML-methods which result in a reduction of the total energy consumption and thus, a prolonged battery lifetime facilitating the coverage of the small intestine, is essential.

Chapter 3

Efficient Machine Learning Approaches Targeting Low-Power Medical Devices

In the field of machine learning, designing increasingly deep and progressively more complex neural networks is a common approach. However, for low-power in-body edge AI devices, such large networks are often not suitable for deployment, e.g. due to their large size with millions to billions of parameters. Thus, applications involving tiny edge devices, which are intended to be equipped with machine learning models to incorporate new functionalities, require hardware-suitable models. For many medical applications, including the neuroimplant and the VCE, most of the published research focuses on achieving proficient classification performance, without considering the possibility to deploy the models subsequently on hardware. This dissertation aims to address this problem by bridging the gap between the current machine learning baseline models and its suitability for succeeding optimization and further hardware deployment.

On the basis of neuroimplants for seizure detection and the VCE for the screening of the small intestine, hardware suitable classifiers should be provided based on current baselines in this field for computer vision and anomaly detection. To accomplish this, for all approaches, the model complexity and overall number of parameters are constantly constrained. Furthermore, the number of needed MAC operations is restricted and either a theoretical realization in hardware of such models is kept in mind or actual hardware simulations are conducted. This includes the consideration of adequate operations, which can be executed by most hardware architectures with minimal adjustments. More specifically, while many operations from neural networks are fundamentally matrix multiplications, many commonly used networks, such as transformers, leverage attention modules, new non-linear functions and operations, which are more difficult to transfer to hardware. For example, the Softmax operation is commonly used in neural networks to compute the class probabilities; however, computing such exponential functions is usually not naturally supported by hardware architectures and therefore, its usage is largely avoided in this work. The basics and challenges of these applications have already been introduced in Chapter 1 and the fundamentals been presented in Chapter 2. This chapter presents hardware-aware machine learning methods, validated on two important medical applications, the neuroimplant and the VCE with the objective to improve upon current baselines while providing hardware-suitable models.

This work aims to generate hardware-aware machine learning models, targeting edge devices, such as the capsule and the neuroimplant, to step towards efficient AI-based low-power medical devices. By incorporating AI, smart decision-making can be enclosed to

lower the overall energy demand and prolong the battery life of such devices. For both applications, in addition to exploring various ML approaches, first results on hardware simulation and corresponding energy consumption results are obtained and discussed. This chapter encompasses content of five peer-reviewed publications that are partially embedded within the individual sections and which can be found in full length in the Appendix A.

In Section 3.1 the EEG-based seizure detection using neuroimplants is discussed, by presenting the publication [Wer+24] (Appendix A1), which targets seizure detection on low-power hardware accelerators. In this work, the time-series properties of the EEG data were leveraged by exploring different time-series analysis methods in combination with neural networks. The whole pipeline was validated and simulated on the ultra-low power hardware accelerator UltraTrail [Ber+20] and the power consumption determined. This is followed by additional results in Section 3.1.2, addressing the model selection process, the possibility to reduce the number of EEG channels to lower the computational demand along with potential extensions of the proposed approach.
Targeting the VCE, Section 3.2 starts with the anatomical classification of VCE images in Section 3.2.1, followed by anomaly detection using an ensemble model in Section 3.2.2 and a multi-task learning approach combining both tasks in Section 3.2.3. This was published in three peer-reviewed publications [Wer+23; Wer+25a; Wer+25b] (see Appendix A2, A3, and A4), respectively. Furthermore, Section 3.2.4 presents an approach to clean VCE datasets from potentially noisy data samples in order to enhance the anomaly detection results. The corresponding publication is attached (see Appendix A5). In addition to the experiments conducted in [Wer+23], in Section 3.2.1 the utilized neural network for the localization task is tested as a quantized model leveraging quantization-aware training and a reduction of the input feature map explored to prepare for deployment. Furthermore, well-established anomaly detection approaches from other fields are presented and transferred to the VCE task to test an additional enhancement of detecting pathologies based on low resolution VCE images in Section 3.2.2. The methods and results from the anomaly detection and multi-task approaches were further refined, combined and presented as additional results in Section 3.2.3. Within this segment, the purely supervised multi-task approach is combined with unsupervised machine learning techniques, which were similarly used for the pure anomaly detection task with the objective to pick out the most prominent advantages of both methods.
Finally, for on-site capsule localization raw image-based organ detection was simulated in software and hardware, which is presented in Section 3.2.5. Hardware simulations yield more reliable results on the actual applicability of the proposed methods. This was published in [Bau+25] (Appendix A6), which discusses the advantage of employing AI-based models in comparison to the procedure without AI support. Chapter 3 concludes with Section 3.3 by introducing possibilities to include additional tasks within the AI-based decision pipeline, such as visibility assessments, as well as a comprehensive disussion in Section 3.4 with a general view on how the designed approaches impact the given medical applications.

3.1 Seizure Detection with Neuroimplants

Seizure detection with neuroimplants is a promising treatment alternative to medical oral therapy and surgery for epilepsy patients suffering from reoccurring seizures. Equipping neuroimplants with an AI-based model allows classification of recorded EEG-signals to perform seizure detection on-site. Upon seizure detection, electrical stimulation is executed, to balance the increased brain activity and limit occurrences of seizures. This potentially

supersedes approaches that constantly stimulate the onset seizure region and lowers the energy demand of such low-power devices. However, due to the limited resources of such tiny medical in-body edge devices, adding additionally functionalities requires careful consideration. For example, when choosing adequate classification models for low-power architectures, their execution necessitates a minimal energy consumption and on-chip area. These aspects need to be addressed while offering real-time classification capabilities to fulfil the purpose of seizure detection before a patient suffers from its consequences.

In the following, state-of-the-art research in the field of EEG-based seizure detection is first outlined and potential limitations are introduced. Subsequently, the peer-reviewed publication [Wer+24] targeting this problem is presented in Section 3.1.1. This provides a pipeline capable of energy-efficient seizure detection with an ultra-low power consumption that is suitable for low-power hardware architectures, such as neuroimplants, and is validated using a commonly used EEG dataset. Finally, simulation is performed using the ultra-low power hardware accelerator UltraTrail [Ber+20]. The manuscript is additionally enclosed in full length in the Appendix A2. In Section 3.1.2, additional results in this matter are presented, displaying more background experiments supporting the aforementioned publication and providing initial experimental results for potential future research.

3.1.1 Energy-Efficient Seizure Detection

Publication A1

Title: Energy-Efficient Seizure Detection Suitable for Low-Power Applications
Abstract: People suffering from epilepsy usually endure reoccurring seizures. Notably, approximately 30% of epilepsy patients are not responsive to anti-epileptic drugs and thus, need an alternative treatment option. Besides surgical removal of the seizure's originating brain region, neural implants are a promising approach to treat arising seizures upon detection using electric stimulation, which balances the increased electrical brain activity and therefore suppresses a seizure. However, these low-power medical edge devices are limited in size and battery lifetime, which call for adequate seizure detection models. Considering these restrictions along with the time-series properties of 1D-EEG data, the lightweight TC-ResNet models combined with additional post-processing using time-series analysis, are a well-suited approach for this application.
In this work, a hardware-sensitive approach involving classification of EEG data using a lightweight neural network quantized to a word width of 4-bit without the necessity of prior feature extraction and in combination with time-series analysis is presented, yielding an accuracy of 95.28% and a sensitivity of 92.34% on the CHB-MIT dataset. Additionally, this model was simulated with the low-power AI hardware accelerator UltraTrail, outperforming other approaches and hardware architectures with a total power consumption of only 495 nW and superior classification performance. Thus, it is demonstrated that this low-power, hardware-aware classification approach is suitable for real-time seizure detection with neuroimplants, providing a new baseline with respect to the energy demand of such tiny medical edge devices intended for epilepsy patients.

Contribution of Authors

Julia Werner:	Conceptual design and implementation. Main author of this publication.
Bhavya Kohli:	Assistance in implementing the preprocessing.
Paul Palomero Bernardo:	Review and assistance with UltraTrail.
Christoph Gerum:	Review and scientific advise.
Oliver Bringmann:	Supervisor and scientific guidance.

State of the Art

In the past, there have been few endeavors to perform successful EEG-based seizure detection while simultaneously limit the overall model size to ultimately equip tiny edge devices with the proposed classifiers. The majority of published results focus on software aspect only. For example, Shoeb and Guttag [Sho09] originally employed SVMs to detect seizures using the CHB-MIT dataset, but without subsequent hardware simulation. Notably, following research on seizure detection was also typically limited to a pure software evaluation of the proposed models. Furthermore, the models were often designed without considering hardware specifications, leading to various difficulties during deployment. For example, machine learning models are commonly trained using the 32-bit floating point representation, which provides a high precision and dynamic range during training.

However, due to the costly implementation of floating point arithmetic in hardware, this can negatively impact the energy efficiency and computation time. Additionally, hardware architectures targeting low-power edge devices usually require models implemented in fixed point representation [PH16; JV23]. [DK20] conducted seizure detection involving CNNs and HMMs using Viterbi decoding including prior feature extraction but without hardware simulation. Based on these findings, within the presented work, time-series analysis is combined with EEG data processing to enhance the CNN classification performance.

Some authors further conducted hardware evaluation of their proposed approach [Hüg+18; Bah+21; Tru+18; Kir+18; Man+22], yielding a power consumption ranging from $< 40mW$ [Kir+18] to $7uW$ [Man+22]. These hardware simulation experiments are consulted as baseline results in the following. Based on these findings, it was intended to reach the current state-of-the-art classification results by real-time seizure detection without prior feature extraction in combination with time-series analysis leveraging the time-series properties of the given data. In addition, it was aimed to lower the required power consumption of the proposed models validated on the low-power hardware accelerator UltraTrail [Ber+20]. This accelerator comprises a configurable array of processing elements and is characterized by an ultra-low power consumption as well as the ability to efficiently execute 1D neural networks. Thus, it qualifies optimally for the given task.

The main objective in the henceforth presented work, was to create a well-performing seizure detection pipeline, including all necessary steps, starting from EEG data preprocessing and concluding with the final hardware simulation. In contrast to [DK20], it was intended to exclude preceding feature extraction and additionally perform hardware simulation outperforming current baselines. Based on current research, a proficient seizure detection pipeline was designed which yields a quantized hardware-suitable model in fixed-point representation, that can ultimately be efficiently executed on low-power hardware accelerators.

Methodology

Leveraging the time-series properties of the EEG data, a lightweight CNN for initial classification was combined with time-series analysis for post-processing and the power consumption validated on the UltraTrail hardware accelerator to assess the proposed method in comparison to current state-of-the-art models. A model of the TC-ResNet family was employed, which combines Temporal Convolutional Neural Networks (TCN) with Residual Networks (ResNet)s, and provides low complex 1D-CNN, which have successfully been used in sensor-signaling applications in the past [Cho+19]. To save chip area and energy, the inherent capability of CNNs to perform feature extraction is leveraged to omit the necessity of implementing a preceding feature extractor in hardware. Furthermore, considering that EEG data constitutes time-series data, three different kinds of time-series methods were conducted: Simple Moving Average (SMA), Exponentially Weighted Moving Average (EWMA) and HMM with Viterbi decoding, to improve the CNN classifications. Thus, after an incoming EEG fragment of 0.5s was classified by a TC-ResNet4 as either a seizure or a healthy fragment, a series of classifications was fed to one of the time-series analysis methods to compute the final most likely class.

To provide a preferably robust model, first, a patient-unspecific TC-ResNet4 was trained, functioning as a base model for all patients. To also account for individual differences within the EEG data between patients, the base model was further fine-tuned using only patient-specific data. Hence, subsequent research can either be build upon the base model or the patient-specific networks. The networks only differ in the weights, but not in the general network architecture, which facilitates the deployment on one general hardware

accelerator. The experiments were first conducted with a 32-bit floating point model, which aligns with the majority of the state-of-the-art models and then adjusted to a quantized 4-bit fixed point representation for the weights, bias and features to validate its hardware suitability. Subsequently, the PyTorch code was converted to C-code and final evaluation was performed on the accelerator UltraTrail, which involves a configurable array of processing elements along with a distributed memory system for dynamic content re-allocation and is generally well-suited for sensor-signal processing [Ber+20]. The design was synthesized in GlobalFoundries 22FDX® 22 nm FD-SOI technology, executed with 10 inferences per second and a clock frequency of 256 kHz.

Results and Discussion

The results are briefly summarized in the following, starting with a comparison of the TC-ResNet4 with varying precision and in combination with a HMM and Viterbi decoding. Additional experiments were performed using the simple moving average and the exponentially weighted moving average method. However, due to the superior performance of the hybrid approach involving Viterbi decoding compared to the moving average methods, the results presented in the following mainly involve the HMM and Viterbi decoding.

Figure 3.1 shows the influence of a varying word width used for the CNN on the overall classification performance. This demonstrates that the word width can be easily lowered to 4 bit, without strong declines in terms of the Area Under the Curve (AUC) score. Furthermore, the plot displays the performance differences between the base model, the patient-unspecific model and the patient-specific model involving the HMM and retraining. During retraining with patient-specific data the individual normal patient recordings function as personalized baselines and the model and the weights re-adapt to detect outliers. While the retraining notably improves the base model, additional Viterbi decoding further enhances the classification results.

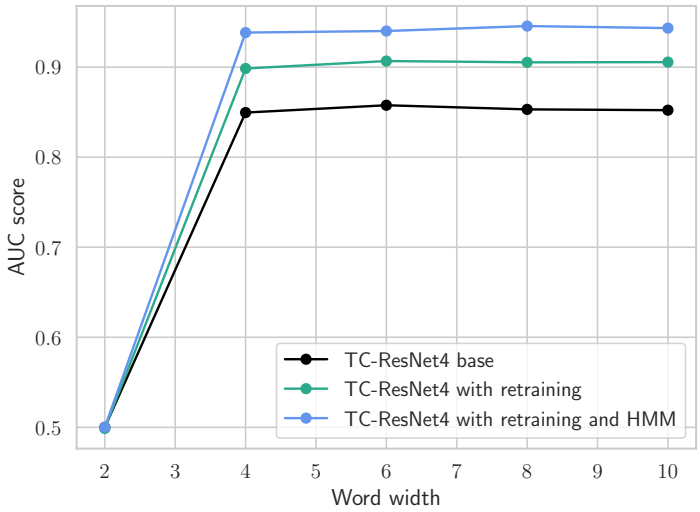


FIGURE 3.1: Classification performance of the base model compared to the patient-specific model and the patient-specific model with Viterbi decoding tested for different word widths, from A1 [Wer+24].

In the following, the classification performance of the 4-bit model with fixed point representation was compared to the TC-ResNet4 with 32-bit floating point representation to

compare how much a reduced precision impairs the classification abilities in this setting (see Table 3.1).

Table 3.1: Comparison of 32-bit vs. 4-bit TC-ResNet4 combined with HMM and Viterbi decoding. Provided metrics in [%].

Model	Accuracy	Sensitivity	Specificity	AUC Score
32-bit floating point & HMM	97.84	92.67	97.93	95.30
4-bit fixed point & HMM	95.28	92.34	95.34	93.84

Comparing the proposed models, the 32-bit approach performs slightly better with an accuracy of 97.84% and a sensitivity of 92.67% than the 4-bit model with an accuracy of 95.28% and a sensitivity of 92.34%. This is expected due to the computation with a higher precision in the 32-bit model. Importantly, since no splits for the CHB-MIT dataset were available, in this work, a new test set was generated as described in the publication [Wer+24]. Although the presented results are in line with the baseline results of [Sho09], who report an accuracy of 96%, direct comparison is difficult due to the absence of a general test set. In this case, slight deviations from the baseline (regardless if better or worse) do not provide major insights on potential improvements. Instead, the main objective was to provide a model that is suitable for low-power hardware architectures and has been demonstrated to perform well with lower precision without significant performance impairment. Notably, the hardware simulation results are independent of the utilized dataset and dataset splits, allowing comparisons across various datasets. While the TC-ResNet4 performed best with 32-bit floating point representation, using a lower precision in combination with quantization-aware training still yields promising results and is additionally suitable for low-power hardware architectures at the same time.

Due to its hardware suitability and remarkable classification performance, the 4-bit model was used for further hardware simulation experiments with the accelerator UltraTrail. Table 3.2 depicts the area cost and power consumption for the individual components of the accelerator.

Table 3.2: Area cost and power consumption of the Ultra-Trail components

Metric	Memory wrapper	OPU	Array	Control unit
Area cost [µm^2]	19,025	1,128	4,504	1,467
Power consumption [nW]	8	66	153	231

While the control unit requires the largest fraction of the overall power consumption, followed by the MAC array, the memory wrapper demands the largest area. The largest share within the memory wrapper is demanded by the weight memory, which stores 9840 parameters for this model. Compared to literature baseline results [Hüg+18; Bah+21; Tru+18; Kir+18; Man+22], the proposed method exhibits the lowest power consumption with only 495 nW and an energy demand of 49.5 µJ. To conclude, the combination of an efficient and lightweight seizure detection pipeline and the usage of an optimized hardware accelerator outperformed current baselines.

3.1.2 Additional Results - Seizure Detection

In the presented work, the TC-ResNet4 was employed as the base neural network architecture for seizure detection. Depending on the number of convolutional layers, this network type can arbitrarily be changed in size, which influences the overall classification performance. Thus, in the following, the impact of the model size on the seizure detection performance is briefly discussed. Within this process, the trade-off between performance and complexity as well as the required chip area and energy demand needs to be considered. Next, the influence of different window sizes for Viterbi decoding on the sensitivity during seizure detection is discussed. Regarding the number of EEG channels, a higher number of channels increases the classification performance, but also naturally raises the computational complexity. Using only a minimal number of EEG channels is desirable, but also potentially lowers the seizure detection accuracy. Therefore, the impact of different numbers of EEG channels on the overall classification abilities of the chosen neural network is inspected. To rule out the possibility that the presented time-series methods only explicitly function well in combination with the chosen network architecture, another well established neural network used for seizure detection in the literature replaces the TC-ResNet and is used for seizure detection. Then, the time-series experiments are repeated and a direct comparison of prediction with and without time-series analysis are conducted. Lastly, a novel approach to extend the current HMM implementation for this setting is outlined to provide insights on possible future directions.

Model Selection and Threshold Determination for Seizure Detection

As discussed in the publication A1 [Wer+24], TC-ResNets qualify well for the analysis of 1D data and additionally, can be efficiently deployed on edge devices. While a smaller network tends to be more suitable for low-power hardware applications, such as seizure detection, the trade-off between classification performance and model size needs to be considered. Thus, for the model selection, the question was raised how different TC-ResNet model sizes impact the overall classification ability. Therefore, the non-quantized TC-ResNet4, TC-ResNet8 and TC-ResNet16, which mainly differ in the number of convolutional layers, were trained, retrained and tested on the CHB-MIT dataset with 16 channels of EEG data and the results presented in Table 3.3.

Table 3.3: Comparison of TC-ResNet4, TC-ResNet8, TC-ResNet16 without additional time-series analysis

Metric	**TC-ResNet4**	**TC-ResNet8**	**TC-ResNet16**
Accuracy [%]	92.46	91.37	93.60
Sensitivity [%]	90.30	88.48	85.66
AUC Score [%]	91.40	89.95	89.71
FPR per h	0.0750	0.0858	0.0625

Interestingly, the largest network, the TC-ResNet16, does not show the strongest performance overall. A possible reason for this is that the problem at hand is fairly simple and choosing a rather complex model does not benefit the objective. It might further be caused simply by noise or a disadvantageous choice of hyperparameters during model training. Overall, the results indicate that the smallest network (TC-ResNet4) has the best trade-off between sensitivity (90.30%), accuracy (92.46%) and model size ($\approx$ 9000 weights). Hence,

the TC-ResNet4 was selected for all subsequent seizure detection experiments and was also employed in the publication A1 [Wer+24].

To perform a standard binary classification as needed for the seizure detection, a default threshold of $t = 0.5$ is usually applied to the last logit θ of a neural network, resulting in class $c = 0$, if $\theta \leq 0.5$ and $c = 1$, if $\theta > 0.5$. However, for imbalanced datasets, simply applying this default threshold can negatively impact the overall classification results. Specifically for classification tasks such as seizure detection, a higher sensitivity at the cost of a reduced specificity is desired, since the main objective is the detection of a large fraction of seizures and a false positive classification has little consequences. Therefore, in the publication [Wer+24], threshold moving was applied to find an optimized classification threshold. Hence, different weighted factors $w \in \{2, 3, 4, 5\}$ were multiplied with the logit specifying the seizure and then, the sensitivity of the model on the train set inspected as shown in Table 3.4.

Table 3.4: Intermediate Results after Threshold Moving with a TC-ResNet4

Weighting applied	**1**	**2**	**3**	**4**	**5**
Sensitivity [%]	86.28	**91.59**	93.99	95.40	96.31

Importantly, due to an absence of concrete guidelines on threshold selection, for this work, the sensitivity was used as a key metric since it is one of the most important metrics for this application. The weight, for which the sensitivity initially surpassed 90%, was selected for all following experiments, which was true for $w = 2$. The threshold of 90% was manually chosen since it was aimed to detect at least 90% of seizure samples. Considering that each sample corresponds to only 0.5s, in reality, the detection rate should be even higher, if a larger time frame of EEG data is considered. Based on these results, in the following experiments, each neural network prediction was weighted with a bias of 2 towards detecting a seizure to enhance the overall sensitivity at the cost of a decreased specificity.

Window Size Definition for Time-Series Analysis

After completing the ML model selection and training, subsequent time-series analysis was applied. An important parameter, which needs to be defined in this matter, is the window size. While a larger window size has the advantage of an enhanced classification performance due to considering a larger time frame of data, it additionally is accompanied by a certain delay and possibly a larger number of values, which need to be stored on hardware. To determine a reasonable window size for subsequent time-series analysis methods, the data from two randomly selected patients, was harnessed to investigate the impact of different window sizes on the classification ability of the TC-ResNet4 while considering the aforementioned factors. Specifically only the HMM in combination with Viterbi decoding was used to determine the window size as it has no additional hyperparameters in contrast to the smoothing average methods.

The results are listed in Table 3.5, which demonstrate that the classification becomes more precise as the window size increases, as expected.

Table 3.5: Impact of different window sizes on the classification performance demonstrated with a TC-ResNet4 and Viterbi decoding.

	2	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
AUC-Score [%]	97.62	97.90	98.42	98.68	98.95	99.21	99.49	99.49	99.50

Notably, an increased window size also correlates with a raised detection delay. Hence, it is advisable to choose a window size which captures differences within the data but is small enough to have only a minimal detection delay. A reasonable choice in this matter can be a size of 5, as it considers the last $5 \cdot 0.5\,\mathrm{s} = 2.5$ s of EEG data leading to a high AUC-score, but is still limited in the detection delay. However, there are no concrete guidelines on how early a seizure needs to be detected so that the stimulation is still on time before any epileptic symptoms appear.

Impact of the EEG Channel Number on Classification Performance and Testing Generalizability

For EEG data classification, the Input Feature Map (IFM) of the TC-ResNet is defined by the number of EEG channels N and the number of used data points K (IFM $= N \cdot K$). The size of the IFM influences the required chip area and therefore should be limited, if possible. Fewer channel result in a smaller first layer of the neural network and subsequently, a reduced memory. Furthermore, in reality, neuroimplants probably cannot offer an arbitrary number of channels by that alone that more channels directly require a larger surgical intervention to implant the electrodes. This motivates the reduction of the total number of utilized channels in conducted experiments. However, a channel number reduction also impacts the classification performance, since a shrinkage in data acquisition results in less information and often a declined prediction performance. One has to consider the trade-off between the channel number and the classification performance. Hence, the impact of using a reduced number of channels for seizure detection was briefly explored by generating three different datasets, consisting of $4, 8$ and 16 channels based on the highest variance. Subsequently, the TC-ResNet4 was trained and retrained on these datasets, time-series analysis applied and finally the AUC score computed. As shown in Table 3.6, with an increasing number of channels, the classification accuracy increases.

Table 3.6: AUC score [%] for different numbers of channels for the non-quantized TC-ResNet4 with SMA, EWMA and Viterbi.

	CNN	**CNN & SMA**	**CNN & EWMA**	**CNN & HMM**
4 channels	87.94	90.92	91.34	92.85
8 channels	90.58	92.68	93.30	94.72
16 channels	**91.40**	**94.08**	**93.94**	**95.30**

The best results are obtained with 16 channels, which has been used in the publication [Wer+24] as well to capture a preferably broad spectrum of data points. It is further demonstrated that a total of 4 or 8 channels still yield a proficient performance, especially in combination with time-series analysis, with an AUC score succeeding 90% in both cases. Thus, using only 8 or even 4 channels is also a reasonable choice and in the future, this

might function as a starting point to further reduce the IFM, depending on the hardware requirements and clinical guidelines on the sensitivity and detection rate.

Actual neuroimplants process the information from a minimal number of channels [Sch+23; Sch+25]. For example, the EASEE® neuroimplant relies on a five-contact electrode in a pseudo-Laplacian arrangement, with the focus area selectively chosen based on the most prevalent onset seizure region [Sch+25]. Consequently, such neuroimplants require accurate seizure detection using only few channels. Notably, the implant presented by [Sch+25] circumvented the use of any seizure detection method, by stimulating every 2 minutes. Equipping such devices in the long-term with a seizure detection algorithm is desirable, since this can lower the stimulation frequency, directly resulting in a prolonged battery lifetime. Hence, in the following, the approach presented in Section 3.1.1 [Wer+24] is additionally tested while including only two channels which are theoretically supported by the EASEE® implant, as shown in Table 3.7.

Table 3.7: Results of the non-quantized TC-ResNet4 compared to the TC-ResNet8 quantized to 6 bit using only 2 channels for data acquisition.

	TC-ResNet4 32-bit		**TC-ResNet8 6-bit**	
	CNN only	& HMM	CNN only	& HMM
Accuracy [%]	88.09	94.57	92.41	96.61
Sensitivity [%]	90.80	92.23	76.77	77.94
Specificity [%]	88.04	94.61	92.71	96.96
FPR per h	0.1196	0.0539	0.0729	0.0304
AUC Score [%]	89.42	93.42	84.74	87.45

This demonstrates sufficient classification abilities of the TC-ResNet4 with an AUC score of 89.42%, which is even slightly higher compared to using 4 channels with 87.94% (see Table 3.6), which were chosen based on the highest variance. Particularly in combination with the Viterbi decoding, the classification performance is improved. However, quantizing the TC-ResNet4 to a word width of only 4 bit in combination with the HMM and Viterbi decoding results in a sensitivity of only 71.69% (results not shown here). Thus, if quantized and restricted to only two channels, the overall model complexity seems to be too low, such that it cannot capture the underlying data structure sufficiently.

Therefore, a larger network (TC-ResNet8) was trained instead and quantized for this setting. Results of the non-quantized TC-ResNet4 compared to the quantized TC-ResNet8 trained on this dataset are also shown in Table 3.7. Quantizing the TC-ResNet8 to 6 bit leads to an AUC-score of 87.45%, an accuracy of 96.61% and a sensitivity of 77.94%. While the performance is still inferior compared to the floating point model, it can be concluded that even with the usage of only 2 channels, the model provides sufficient classification abilities and can function as a starting point for further optimization.

It was generally demonstrated that the combination of CNN prediction and time-series analysis with Viterbi decoding is a well-suited approach for seizure detection. To exclude the possibility that this is network dependent, the effectiveness of subsequent time-series analysis was additionally tested in combination with a well-established neural network proposed by [Man+22], which is a simple CNN consisting of three 2D convolutional layers. Using this model, the training pipeline was precisely conducted as before and the classification performance including time-series analysis compared to the original results without such post-processing. Before subsequent time-series analysis, an AUC score of 88.59% was achieved on the CHB-MIT dataset with this model, which is concurrent with

the reported results from the authors [Man+22], who yielded a mean AUC score of 87.6% on the CHB-MIT dataset with 4 channels. Small differences can occur due to a different split, preprocessing and different channel selection compared to [Man+22]. Table 3.8 lists the results with additional post-processing and shows that smoothing by different simple time-series methods yields a significant improvement with an AUC score enhancement up to 5.48 percentage points.

Table 3.8: Results of the CNN (quantized to 8 bit) proposed by [Man+22] combined with the presented time-series analysis methods.

	Moving Average	**Exp. Moving Average**	**HMM**
Accuracy [%]	95.87	94.78	93.97
Sensitivity [%]	88.72	90.15	94.16
Specificity [%]	96.00	94.87	93.97
FPR per h	0.0400	0.0513	0.0603
AUC Score[%]	92.36	92.51	94.07

The proposed 2D convolutional CNN of [Man+22] has a similar number of parameters (10,145 weights) as the TC-ResNet (9,840 weights) which was used in the experiments in this work. Considering that some hardware accelerators might only be able to execute 1D but not 2-Dimensional (2D) networks, it was focused on a network consisting only of 1D convolutional layers, which is given by the TC-ResNet family. Furthermore, in the publication A1 [Wer+24], a total energy demand of 0.0495µJ with the UltraTrail hardware accelerator was reported using the TC-ResNet4. This outperforms the findings from [Man+22], who used an ultra low-power Apollo 3 Blue microcontroller with a total energy demand of 7.01µJ. This underlines the finding that applying subsequent time-series analysis is a valuable approach in this setting relatively independent of the network architecture, which not only improves the classification abilities, but presents also a hardware-suitable approach.

Extension for Accuracy Enhancement of the Viterbi decoding

In the current hybrid approach of a CNN and subsequent Viterbi decoding based on a HMM, the predictions of the neural network are directly passed to the HMM. In a binary setting, the neural network node obtaining higher values, simply corresponds to the most probable class. However, this is accompanied by an information loss, that can theoretically be prevented to some degree. One possibility is to store the evaluations of the CNN more precisely instead of forwarding the final CNN prediction, which is outlined in the following.

Originally, in the work presented in Section 3.1.1 [Wer+24], the HMM model was designed with two hidden states: s_{normal}, in the case of normal EEG data, and s_{seizure}, if a seizure occurred. Emissions can be either $e = 1$, if the output neuron for seizure θ_{seizure} displays a higher value than $\frac{1}{2}$ and $e = 0$, if the normal state θ_{normal} presents a higher value than $\frac{1}{2}$.

However, propagating only this information does not consider the model's confidence on the predictions and thus, is accompanied by an information loss. To counteract this phenomenon, one alternative approach is to introduce a larger number of possible emissions. For example, additional emissions could correspond to *seizure*, *most likely a seizure*, *most likely no seizure* and *no seizure*, depending on the output probability of the neural network.

Following this, the described setting is arbitrarily extended to a larger number of emissions. This is depicted in more detail in Figure 3.2.

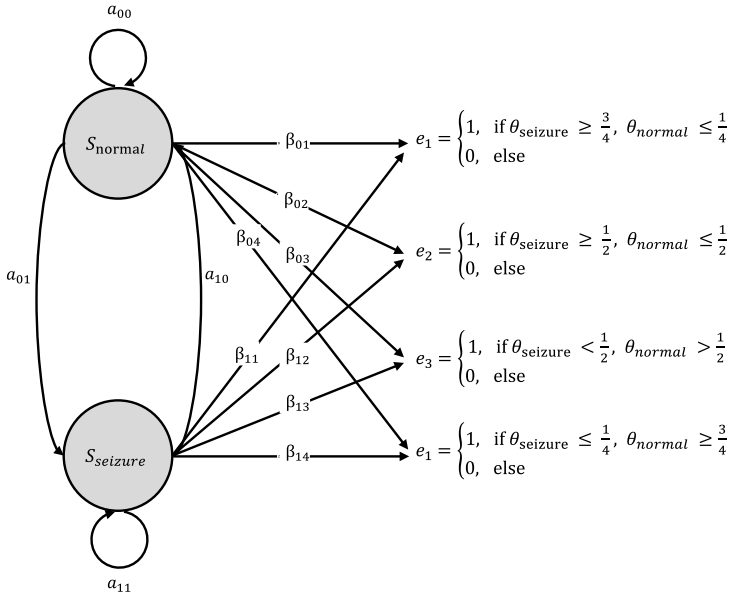


FIGURE 3.2: HMM overview with two hidden states and the four possible emissions e_1, e_2, e_3 and e_4.

Theoretically, an increased number of emissions results in a more precise information retrieval. Nevertheless, this is also accompanied by an increased computational workload and demands a larger area for storing the extended matrices. Resulting from this new setting, there are different research questions which arise. For example, the determination of the threshold for the class decision can be modified to increase the model's sensitivity. In Figure 3.2, the threshold was evenly set, but it might be beneficial to consider additional factors, such as probability distribution of the different emissions. Furthermore, this parameter could be incorporated into a NAS in order to find a better threshold of emissions for a given neural network.

Besides extending the HMM, introducing a third class in addition to *seizure* and *normal* might improve this approach and enhance the classification abilities. Current research indicates that before a seizure starts, the brain activity already shows an abnormal behavior. To address this, a third class can be useful to correspond to an intermediate label between seizure and normal. However, the reported time frames vary heavily with findings from a few seconds [PDG13], over several minutes [MC14], to 2-7 h [Lit+01], since physicians disagree on a clear starting point of a seizure and a general time period in which differences are observable before a seizure occurs. This is further complicated by individual differences between patients. Since the CHB-MIT dataset only offers the healthy and abnormal classes, the dataset was manually adjusted, such that a third class was added, which was labeled as *intermediate*, comprising the last 20 seconds before a seizure occurred. Nevertheless, training a TC-ResNet with this modified dataset and combining it with the HMM, yielded inferior results compared to the original database. One possible reason for this outcome is the very little amount of available data in this class, which was inherently capped by the total number of seizures. The sparse data might not be sufficient to train the model on those three classes, leading to strong underrepresentation of the smallest class. Therefore, the experiments could be repeated on a larger dataset, multiple databases pooled together or tested with a larger period of interictal data. While for some patients a specific time-period still corresponds to a striking preictal data period, for others this only represents normal

EEG data. Thus, the chosen time period should not be too large and must be individually adapted. Based on the aforementioned findings, within this work, a promising seizure detection pipeline targeting low-power neuroimplants was presented.

3.2 Enhancing Video Capsule Endoscopy - Anatomical Classification and Anomaly Detection

The Video Capsule Endoscopy is another medical procedure which can benefit from machine learning models tailored for this use case. For this application, this dissertation has four main objectives, which are depicted in Figure 3.3 and explained further below.

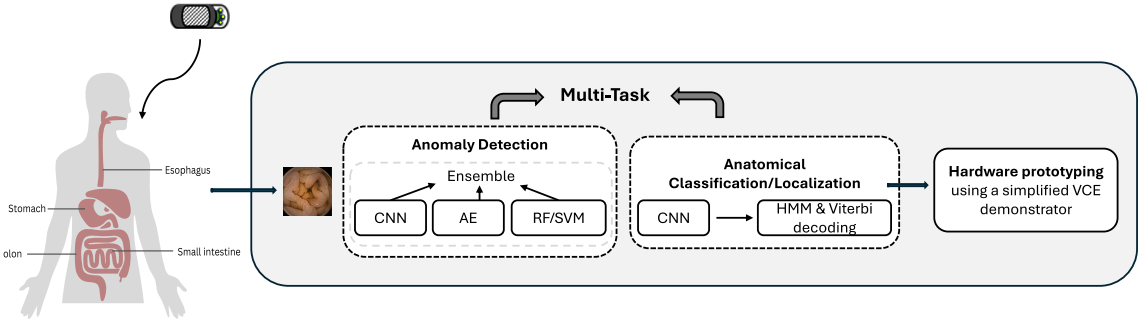


FIGURE 3.3: Main objectives for the VCE, which are addressed in this dissertation: localization (organ classification), anomaly detection, designing of multi-task approaches to combine both and first hardware prototyping with a simplified network.

As a basis to fulfil these tasks, the available data was gathered from datasets consisting only of VCE data from real medical patient VCE studies, covering the GI tract. First, it was aimed to perform precise localization based on VCE images, by organ classification within the GI tract to determine accurately when the small intestine was entered. It is expected that the precise stomach exit identification saves energy by omitting the necessity of transmitting images beforehand. In this work, localization refers to the anatomical classification over time rather than the involvement of segmentation as may be common in other fields. The anatomical classification is realized by using a CNN in combination with a HMM and Viterbi decoding as discussed in Section 3.2.1.
Second, anomaly detection should be performed leveraging anomaly detection techniques from other fields, which allows adding new functionalities to this procedure upon finding a pathology. This is supposed to be utilized after the capsule has entered the small intestine and is presented in detail in Section 3.2.2.
Third, both preceding tasks should be combined in a single multi-task model while competing with current baselines (see Section 3.2.3). This theoretically allows to employ a model capable of localization as well as anomaly detection. To address the anomaly detection task from another perspective, Section 3.2.4 discusses the effectiveness of cleaning VCE datasets from noisy labels before conducting image classification. The anomaly detection results obtained using a cleaned training set are then compared to preceding results.
Lastly, by employing a lightweight CNN, the localization task should be tested using a simplified hardware demonstrator simulating the VCE setting, which is presented in Section 3.2.5. It is aimed to reduce the total energy demand of the AI-based approach compared to current standards in literature.

3.2.1 Localization within the GI Tract (Publication A2)

Publication A2

Title: Precise Localization within the GI Tract by Combining Classification of CNNs and Time-Series Analysis of HMMs

Abstract: Organ classification within the GI tract based on VCE images assists in localizing the capsule and identifying when the organ of interest, the small intestine, is entered. This paper leverages the image classification properties of CNNs and the knowledge of the natural order of organs that the capsule traverses by mirroring this in the transition probabilities of a HMM. By combining a lightweight CNN with the time-series properties of a HMM and subsequent Viterbi Decoding, to compute the most likely sequence of organ states, the presented approach outperforms current baselines while additionally limiting the overall model size. Validated on the Rhode Island dataset, the largest publicly available dataset including organ labels, our approach achieves an accuracy of 98%, outperforming current baselines. This can function as a starting point for further optimization.

Contribution of Authors

Julia Werner:	Conceptual design and implementation. Main author of this publication.
Christoph Gerum:	Scientific advise and proofreading.
Moritz Reiber:	Scientific advise and proofreading.
Jörg Nick:	Scientific advise and proofreading.
Oliver Bringmann:	Supervisor and scientific guidance.

State of the Art

Determination of the capsule's localization within the GI tract has been addressed with approaches such as magnetic tracking of the capsule in vivo [PA14; YS13; Wan+21; FG22], video-based estimation of the capsule by measuring the displacement and rotation between frames [ZBP14], or even using digital radiography [Mar+14], in which after a defined time has passed, the abdomen is captured by X-rays. Furthermore, popular approaches involved radio frequency [Fis+04; BMP13; Ye+12; Pah+12] and analyzed the received signal strength. Additionally, image-based methods relying on VCE frames to estimate the capsule's movement and speed evolved [Bao+14; Bao+12; Iak+13]. Some of these approaches require external sensors or devices to help track the capsule. In contrast to that, purely image-based methods usually only rely on the capsule itself.

In the present work, the main objective was to perform precise localization within the GI tract relying only on VCE images. Instead of constantly tracking the capsule's movement or speed, the primary goal is to classify the current organ, at which the capsule currently resides, and to define the entering of the small intestine accurately. This allows to transmit images for further evaluation starting only at this moment and to discard all other images until then, which do not capture the region of interest. Limiting the costly transmission of images until the small intestine is reached reduces the energy consumption of the overall procedure since fewer images need to be delivered in total. In the past, there have been

further endeavors to improve the anatomical detection. For example, [Lee+07] performed organ classification based on color change patterns and different patterns of intestinal contractions. Later, [Zou+15; Chu+23] used deep neural networks for organ detection based on VCE images, however without focusing on the issue of hardware suitability and by using non-public datasets, which restricts the reproducibility. Leveraging the release of the largest labeled VCE dataset involving organ annotations, the approach was validated on the Rhode Island dataset [Cha+22], published in 2022. The current baseline for this data is provided by [Cha+22], who report an accuracy of 97.1% using an Inception ResNet V2 with a total of 56 Million (M) parameters. When targeting low-power devices for this application, it is necessary to reduce the given neural network complexity. Thus, it was aimed to reduce the overall model size while obtaining a comparable accuracy as the current given baseline.

Approach

A large portion of the capsule's energy is needed for transmitting images to an external device. Additionally, simply storing all images locally on the capsule does not enhance the battery lifetime or allow real-time assessment on-site and further requires retrieval of the capsule after the procedure [Zwi+19]. Importantly, only images of the small intestine are of interest for further inspection. Therefore, starting the image transmission only directly after the small intestine was entered, can save energy since images of the preceding and not relevant organs are not sent out anymore. To precisely determine the location of the capsule on-site, combining the classification abilities of neural networks with the time-series analysis properties of HMMs was conceptualized, as illustrated in Figure 3.4.

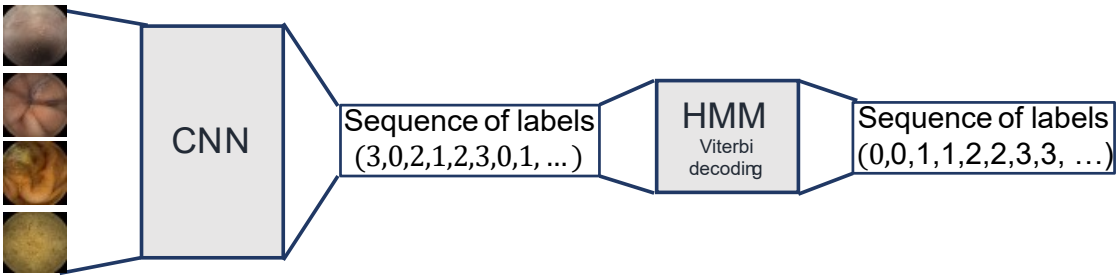


Figure 3.4: VCE Image processing pipeline using a CNN, a HMM and Viterbi decoding, from [Wer+23].

As described in Chapter 2, the MobileNet is a reasonable choice as a lightweight CNN and was implemented for the following experiments as the neural network of choice. After receiving VCE images sequentially, a class for each frame is returned by the CNN and then passed to the HMM, which subsequently computes the most likely sequence of organ labels with Viterbi decoding. This leverages the fact that the VCE data was captured and is available in chronological order. Hence, this can be considered a time-series problem, for which the HMM and Viterbi decoding are well-suited.

While computing the most likely sequence of states with Viterbi decoding, a log-likelihood matrix is built, storing the possible paths of sequences. However, when targeting tiny in-body edge devices as needed for the VCE, every part of a classification approach should be limited in computational complexity to restrict the required chip area for each pipeline segment. Thus, to limit the size of the log-likelihood matrix, a sliding window was implemented, which defines the number of classification labels from VCE input images, that are considered for each computation. Given the inherent anatomy of the GI tract, the order of states/organs, which are traversed by the capsule, are known. Additionally,

it is known that the capsule cannot skip organs or jump back to a previously traversed organ due to the human anatomy. This knowledge is directly encoded in the transition probabilities to improve the evaluation performance. Furthermore, the CNN predictions retrieved from the training data and the resulting confusion matrix directly encode the emission probabilities of the HMM and can thus be employed during subsequent Viterbi decoding. Based on this prior knowledge, the presented approach is expected to filter the CNN output and to enhance its predictions leading to an improved localization within the GI tract.

Results and Discussion

In the following, the organ classification results with the proposed approach in comparison to the baselines are presented, followed by an overview of two example VCE studies. First, the main results are presented in Table 3.9 compared to the current baseline.

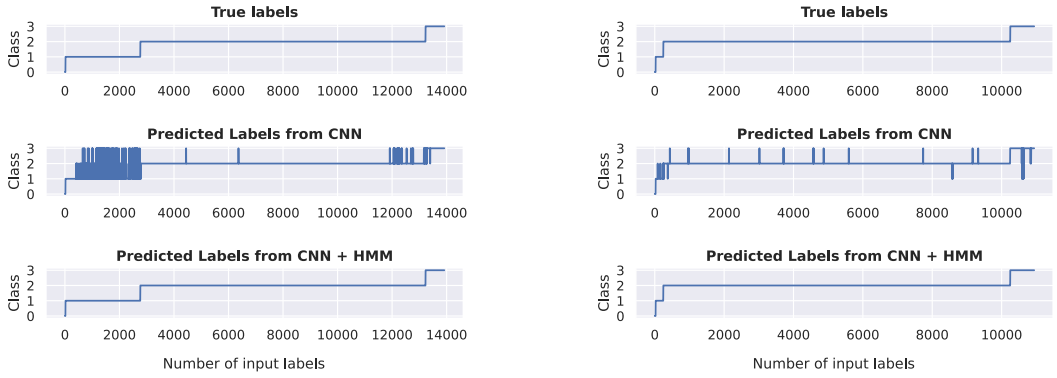
Table 3.9: Results for organ classification by combining a CNN with a HMM compared to the baseline.

Model	**Accuracy [%]**	**# Params**	**Delay [Frames]**
Inception ResNetV2 [Cha+22]	97.1	56 M	-
MobileNetV3	96.95	1 M	-
MobileNetV3 with HMM	**98.04**	**1 M**	19

The MobileNetV3 in combination with the HMM and Viterbi decoding outperformed the CNN on its own with a total accuracy of 98.04% vs. 96.95%. In comparison to the baseline [Cha+22], the classification performance is not only improved, but additionally the model size reduced from $\approx$ 56 M parameters to $\approx$ 1 M parameters. The obtained model size reduction in combination with an outstanding classification performance presents an important step towards hardware suitability. One drawback of the presented method is that the organ detection is accompanied by a certain delay, corresponding to the processing time of 19 frames on average. Using an increased window size during Viterbi decoding correlated positively with a boosted accuracy, but also with an increased delay of the small intestine detection. The CNN on its own does not possess such delay besides the general inference time. However, since the first part of the small intestine can usually be assessed by prior gastroscopy, a small detection delay of the stomach exit does not constitute a major impediment.

In the publication A2 [Wer+23], for the patient studies with the ID 94 and 95, the classification performance was more closely investigated. Similar observations were made almost across all VCE studies and are exemplarily shown for those two patients in Figure 3.5, which displays for each VCE image ordered over time, the true label on top, the predictions from the CNN in the middle row and finally, the evaluations of the CNN and HMM combination at the bottom row.

Besides the classification improvement obtained by the CNN and HMM, the hybrid approach also enhanced the image detection abilities structurally, by smoothing wrong predictions from the CNN based on statistical prior knowledge on the GI tract. Since the defined transition and emission probabilities do not allow jumping backwards to a previously detected organ, once the small intestine is detected by the Viterbi decoding, no classification of the esophagus or stomach is considered valid by the algorithm. This results in fewer misclassifications while the capsule moves forward. Patient studies with



(A) VCE study ID 94.

(B) VCE study ID 95.

FIGURE 3.5: Classification performance comparisons of different approaches exemplarily shown for two patients, from A2 [Wer+23]

inferior results were investigated more closely to find a reasoning for the observation of these outliers. Mainly, if the CNN already performed poorly, the HMM is not performing well either, since it is built to trust the CNN predictions to some degree via the emission probabilities. Poor performance of the CNN on single patient studies can be caused by a bad image quality, e.g. due to digestion remains or air bubbles within the GI tract.
Considering the limitations of such VCE devices, reducing the computational complexity must be addressed further. One possible approach is to reduce the input size of the images of interest. To explore this, images were downsized from 320×320 to 64×64 to reduce the size of the large first layer of the neural network. This resulted in an only marginally impaired performance compared to the original input image sizes, with a final accuracy of 96% and demonstrated that this is a promising approach to further limit the complexity (detailed results can be found in the Publication A2). In summary, combining the time-series properties of the HMM and the vision capabilities of neural networks results in an improved organ detection and localization within the GI tract as well as a more accurate detection of entering the small intestine, the main organ of interest during a VCE.

Additional Results

In the following, additional results contributing to the publication presented above, are shown, providing more insights to the model performance and first experiments involving a quantized model considering further hardware suitability.
In the Publication A2 [Wer+23], for each VCE study, the accuracy obtained by the CNN on its own in comparison to the evaluations from the CNN and HMM in combination are presented. On that basis, in Figure 3.6 the same plot is replicated including the F1-Scores for each study, which besides the accuracy score, forms an important metric to evaluate the classification performance of machine learning models as it presents the harmonic mean of precision and sensitivity.

The F1-Score is plotted on the y-axis for each VCE study from the Rhode Island test set as indicated on the x-axis and the evaluations of the CNN as well as from the hybrid approach visualized. Similarly, as seen with the accuracies for each VCE study [Wer+23], the F1-Score is notably enhanced if the hybrid approach is used. For only very few studies, the CNN by itself performs better than the hybrid approach. In the majority of cases, this is true when the CNN itself cannot classify the images very well. The poor confidence of

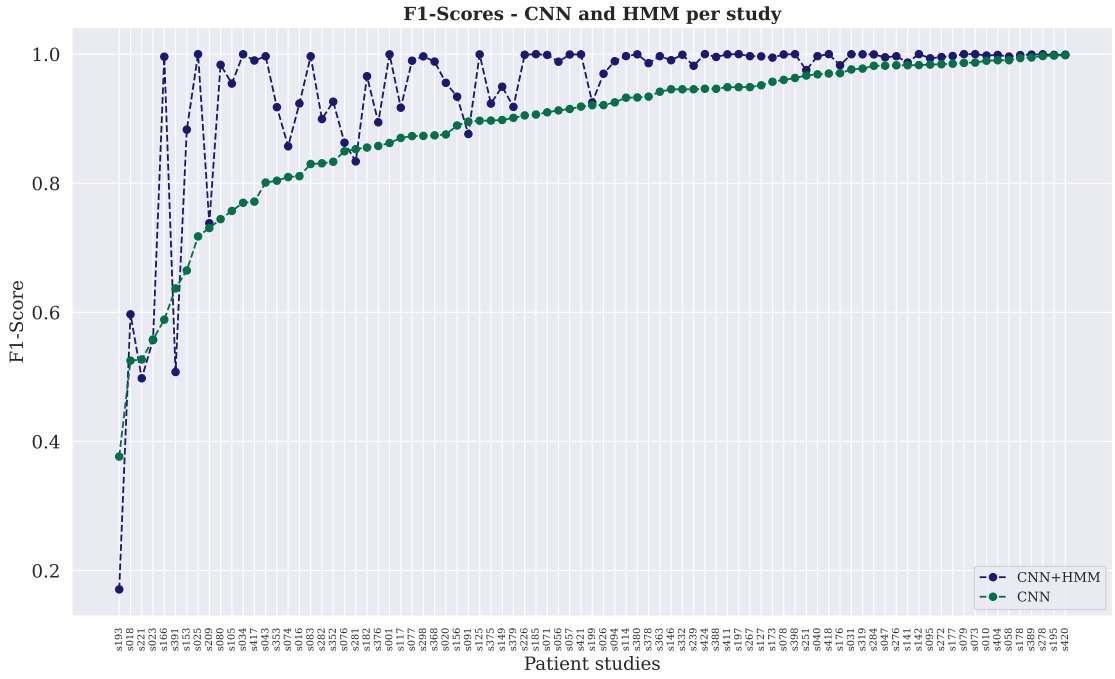
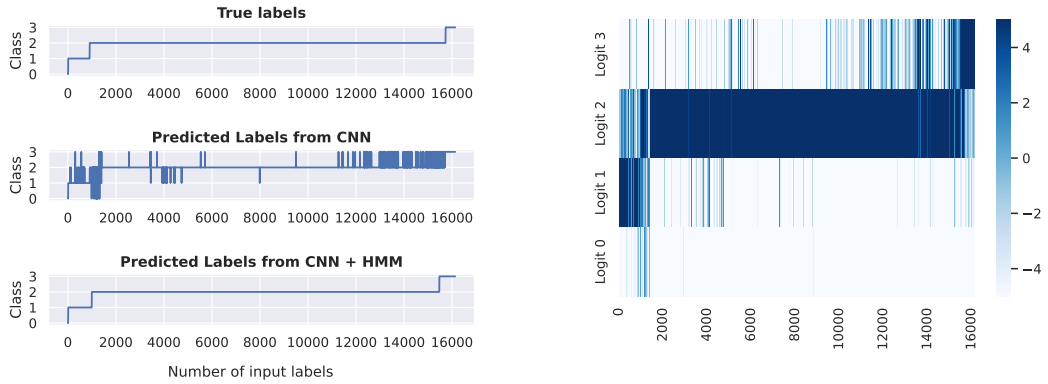


FIGURE 3.6: F1-Scores obtained by the CNN compared to the hybrid approach of CNN and HMM.

the CNN then transfers to the HMM, leading to little or no improvement. Considering the CNN's confidence during Viterbi decoding might improve the overall performance, which is thematized in the next paragraph. A possible reason for a weak performance of the CNN on single VCE studies is potentially a bad video quality due to digestion remains, a dirty camera lens or air bubbles. Overall, the combination of the CNN, HMM and Viterbi decoding outperforms the CNN on its own, achieving F1-Scores of up to 100%.

In the Publication A2 [Wer+23], summarized in Section 3.2.1, the class predictions from the MobileNetV3 as well as the HMM and Viterbi decoding were inspected over time. While this visualizes the classification performance for individual patients very well, this does not incorporate any information on the classifier's confidence. Thus, instead of simply indicating the class, for which the neural network obtained the highest value, the values of all four logits corresponding each to one of the four classes are plotted over time for one patient exemplarily, as shown in Figure 3.7.

This emphasizes that a lower confidence of the CNN correlates with an increased number of false predictions (e.g. for images between 12,000 and 16,000). This can also be observed at the moment of transitions. For ranges of indices in which numerous mislabels occur (e.g. in the strip just before index 14000), the uncertainty of the classifier can be directly observed within the logits. This uncertainty is measurable both by the occurrence of false predictions, which are corrected by the Viterbi filtering, and by the relatively low difference between the two maximal logits. Interestingly, clinicians sometimes also have difficulties in accurately detecting the transition from the small intestine to the colon, which might additionally increase the uncertainty at this point due to potential mislabeling. The additional information provided by the logits is neglected by the presented time-series approach in [Wer+23], which is mainly based on the classifier's output for each moment and acts independently of the numerical value of the logits. Nevertheless, the presented approach involves the classifier's uncertainty indirectly by encoding this information in



(A) Class predictions. (B) CNN logits.

FIGURE 3.7: Classification performance of the CNN compared to the HMM and Viterbi decoding (3.7a) and the CNN logits for all four classes over time (3.7b), exemplarily shown for study s020.

the emission probabilities. Other time-series approaches that are based on filtering time series data, including numerical values at each moment such as recurrent neural network architectures, for instance Long Short-Term Memory (LSTMs), might benefit from the additional information provided by the logits. However, using the logits as likelihoods instead of the discreet class predictions in the combinatorial approach of CNN and Viterbi decoding also possesses shortcomings. For example, all likelihoods for all classes across the whole window size must be stored on-device. Additionally, the emissions currently encode information on the neural network's false predictions and thus, its uncertainty. This is then included in the Viterbi decoding. Directly using the logits would require careful consideration on how to build a new emission matrix in this setting. In addition, a parameter search for the combined HMM classifier might also already implicitly use this uncertainty in the logits, in particular in the context of threshold moving, which can be also investigated in future experiments, for example by involving a NAS.

Furthermore, in the presented publication, only 32-bit floating point models were tested. However, for hardware-aware models, it is essential that the weights are present in quantized, fixed point representation, to limit the power consumption and memory usage. As a consequence in the following, the MobileNetV3 was quantized by employing quantization-aware training. Additionally, the classification performance of a significantly smaller model compared to the MobileNet was explored for this task to provide an initial assessment of the possibility to further lower the model complexity. The MobilenetV3 was fine-tuned with quantization-aware training with 8 bits allocated for the weights, which resulted in an accuracy of 91.07% (see Table 3.10).

TABLE 3.10: Comparison of quantized MobileNetV3 and subsequent Viterbi decoding ($w = 300$).

Metric	**MobileNetV3**	**MobilenetV3 & HMM**
Accuracy [%]	91.07	96.12
MAE	0.1091	0.0620
R2-Score	0.2365	0.6832
Avg. Delay (# Frames)	–	14.86

Subsequent application of the HMM and Viterbi decoding increased the final accuracy

to 96.12%. Compared to the floating point model (98.04%), the accuracy is slightly reduced. However, since an accuracy of 96.12% indicates a proficient performance, it can be concluded that this quantized model is eligible as a starting point for inference on hardware and that specifically in combination with the Viterbi post-processing preserves a good classification performance.
While the utilized MobileNetV3 is already restricted in its size compared to conventional and standard networks, such as transformers, the absolute number of parameters (1 million) still leaves room for improvement. Hence, a significantly smaller convolutional neural network (4 convolutional layers, ReLU activation function, 1 fully connected layer, stride 2, kernel size of 3) was designed and trained on the Rhode Island dataset for the classification of different organ sections in the GI tract. Next, Viterbi decoding was employed with a window size of $w = 300$ for each test study. The network was trained and tested on images with a resolution of 320×320 and with 32×32 in order to further reduce the model size and the needed number of MACs. The corresponding results are presented in Table 3.11.

Table 3.11: Results of a simple CNN with subsequent Viterbi decoding ($w = 300$) and a resolution of 320×320 as well as 32×32.

	320×320		32×32	
Metric	**CNN**	**CNN & HMM**	**CNN**	**CNN & HMM**
Accuracy [%]	88.52	94.53	87.83	92.95
MAE	0.1376	0.0757	0.1327	0.0766
R2-Score	0.0169	0.5672	0.1799	0.5296
Average Delay (# Frames)	–	22.24	–	36.92

The results demonstrate that the applied model for this task can be further reduced in size and although the classification performance is slightly impaired if trained with 32×32 input images, the model trained with a reduced input image size still obtains an accuracy of $> 90\%$ in combination with Viterbi decoding. The final convolutional neural network consists of only $31,576$ weights in total ($17.5M$ MAC operations), with the original image size or $6,232$ weights in total with the lower resolution, respectively. This demonstrates a proficient prediction performance and can function as a starting point for subsequent experiments targeting the reduction of the input feature map and the size of the first layer of a CNN, e.g. by employing a NAS.

3.2.2 Anomaly Detection for VCE (Publication A3)

Publication A3

Title: Enhanced Anomaly Detection for Capsule Endoscopy Using Ensemble Learning Strategies

Abstract: The limited size of a video capsule used for endoscopies restricts the model types and complexity, which can be designed and employed on such a small device. When targeting anomaly detection on-site, this limitation along with challenges due to sparse amounts of data need to be overcome. In this work, anomaly detection based on VCE images is performed by leveraging ensemble methods, while constantly limiting the total number of parameters. The concept of ensemble learning is transferred to this problem to utilize the predictions of various, independently trained models and to finally combine those predictions to achieve a superior result. Potential weaknesses of individual approaches can hereby be balanced leading to an enhanced classification performance. To realize this, multiple loss functions, which are well-established in the field of anomaly detection, were implemented and used to train different models. This approach was tested on the two largest VCE datasets, Galar and Kvasir-Capsule, resulting in an AUC score of $\approx 77\%$. Thus, this work contributes a hardware-aware ensemble learning strategy for anomaly detection tasks, which is ultimately validated on the problem of anomaly detection on VCE images. With only a fraction of memory costs, this approach outperforms state-of-the-art models in this field.

Contribution of Authors

Julia Werner:	Conceptual design and implementation. Main author of this publication.
Christoph Gerum:	Scientific advise and proofreading.
Jörg Nick:	Scientific advise and proofreading.
Maxime Le Floch:	Provision of the dataset and contributing medical insights.
Franz Brinkmann:	Provision of the dataset and contributing medical insights.
Jochen Hampe:	Secondary supervisor.
	Provision of the dataset and contributing medical insights.
Oliver Bringmann:	Supervisor and scientific guidance.

State of the Art

Ensemble learning has been applied in various fields to perform anomaly detection for classification performance enhancement of individual models by leveraging the strengths of multiple methods [Yu+10; ZM12; VC17; ALN20; She+20]. Consequently, this typically results in more complex models than the individual approaches, which potentially limits its applicability for applications involving low-power edge devices.
Furthermore, anomaly detection is of interest in several applications, including the field of medicine [CBK09; Nas+18; Zha+19]. For outlier detection, if large amounts of unlabeled data exist, unsupervised techniques can exploit this advantage, for example, by using

deep autoencoders as demonstrated by [AC15; Gon+19; ZP17; Zon+18]. However, when using ensemble models targeting edge devices, considering the on-chip area and energy constraints of available hardware for a given application still continues to be an ongoing challenge. While this is addressed in the following results, the primary goal in this work is firstly to improve the anomaly detection on VCE images without subsequent hardware simulation.

For video capsule endoscopy, the Kvasir-Capsule [Sme+21] and Galar [Le +25] dataset include annotations for anomalies and are therefore the primary choice for addressing the challenges of anomaly detection in this field. The Galar dataset has just recently been published, such that the current research baseline for this dataset is solely provided by the original authors [Le +25]. While the authors do not perform image classification on all anomalies pooled together, they tackle this problem for each anomaly individually. Since detailed assessment of pathologies on captured images can be performed by trained medical doctors after the images were transmitted and a fine-grained classification of individual anomalies requires a larger model size, the following work aimed to provide an additional baseline for classifying all anomalies in a binary manner (healthy or anomaly). Additionally, compared to the work of [Le +25] the model size was further limited to achieve more hardware-friendly models for this application.

For the Kvasir-Capsule dataset, research has been conducted targeting anomaly detection for all pathological classes [Sme+21; Sri+22; Reg+25]. However, this is difficult to compare to the following work, which conducts a binary study. The most fitting reference for comparison is provided by [Sá+23], who conducted classification with binary classes as well. Regardless, none of the aforementioned studies focused on generating results with hardware-friendly models, but used deep neural networks with many parameters (e.g. the ResNet152 employed by [Sá+23] consisting of 60 M parameters, making it unsuitable for low-power architectures).
In the following work, established techniques from the field of anomaly detection and ensemble learning were combined to progress the anomaly detection based on VCE images, while focusing on models with limited complexity and parameters.

Approach

When targeting anomalies within medical data, potential class imbalances need to be considered to prevent the model from mainly being trained on healthy samples and therefore introducing a bias resulting in a high specificity. To address these omnipresent class disparities in VCE data, first, thorough preprocessing of the datasets needs to be performed and an adequate sampling of the anomaly class conducted. One possible option is to sample only anomalies within the small intestine and disregard anomalies present in other organs to focus on the relevant anomalies only. This was only conducted for the Galar dataset, since the Kvasir-Capsule dataset does not offer organ annotations along with the pathological labels. As a result, the Kvasir-Capsule dataset was directly adapted as originally published. For the Galar dataset, the following most prominent anomalies were included: polyp, blood, active bleeding, angiectasia, erosion, erythema and ulcer. To address the class imbalance and therefore counteract a bias within a trained model, data augmentation along with a weighted sampler was employed, which balances each batch according to the number of labels in each class.

To further balance potential weaknesses of individual models, an ensemble model is used to strengthen the overall predictions. The presented model involves four different methods as shown in Figure 3.8: 1) standard image classification in a supervised manner, 2) unsupervised machine learning with an autoencoder, 3) a semi-supervised model including

an autoencoder and a linear classifier head and 4) a random forest or SVM as the final classifier.

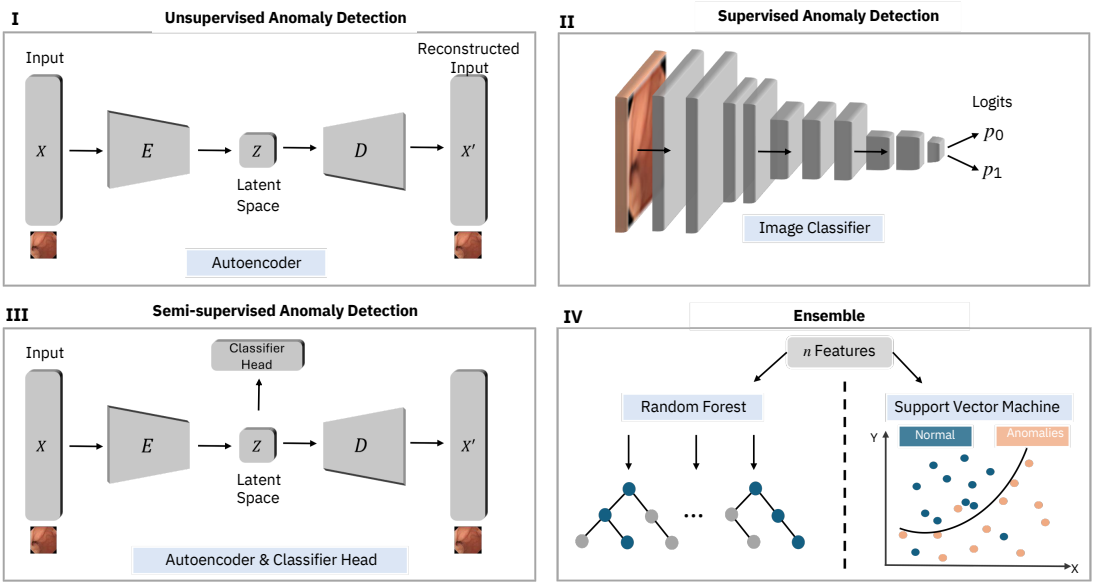


FIGURE 3.8: Depiction of the ensemble learning strategy for anomaly detection, comprising an unsupervised (I), a supervised (II) and a semi-supervised (III) classification approach, concluded with an ensemble model with either a random forest model or a SVM (IV). I, II and III are thereby based on the MobileNetV3 network architecture, from [Wer+25a].

The autoencoder is used to leverage the large amount of healthy data in the Galar dataset and a vast majority of unlabeled data in the Kvasir-Capsule dataset. It is trained to reconstruct a given input image and based on the resulting reconstruction loss, a difference compared to a ground truth is created. The underlying assumption is that if the autoencoder is predominantly trained with healthy samples, normal images obtain a small reconstruction error while pathologies should optimally lead to a larger loss. In theory, this method can perform well, even when only healthy data is present, as long as the difference between healthy and anomaly samples is large enough. This reconstruction error is then employed as a feature for the final ensemble model. As this directly benefits from the existing class imbalances, this is a well-suited approach for the available datasets. Furthermore, within the latent space, the input features are compressed in a lower dimensional space, which is additionally leveraged by adding a small classifier head which itself returns a classification for each input. The output of the classification head is also used as a feature for the ensemble model. Lastly, the evaluations of a standard image classifier are included in the ensemble model. The final ensemble model, consisting either of a SVM or a random forest model, evaluates the predictions of each method and then provides a final prediction. This should ensure an enhanced classification performance in comparison to using only the individual networks.

This approach leverages the vast majority of unlabeled frames in the dataset by using an autoencoder and subsequently finding differences compared to labeled anomalies. Combining this technique with supervised and semi-supervised methods is expected to improve the overall prediction performance. In addition to the enhanced image classification, this approach constantly constrained the model size and the total number of parameters to step towards hardware suitable machine learning models. Furthermore, all different methods described above involve the same backbone architecture of a MobileNetV3 to facilitate

easier deployment, while architectures such as vision transformer are completely neglected due to their infeasibility for low-power architectures.

Results and Discussion

In the following, the results of the ensemble model are discussed in comparison to the state-of-the art models in this field and further analyzed how the invidual approaches contributed to the final prediction. First, in Table 3.12, the best ensemble model results are listed. This can be inspected in more detail in the Publication A3 [Wer+24].

Table 3.12: Results of the ensemble model validated on the Kvasir-Capsule and the Galar dataset.

Dataset	**Model**	**AUC score [%]**
Kvasir-Capsule	Ensemble RF, AE, CLF	76.86
Galar	Ensemble SVM, AE, CLF	76.98

For both datasets, the ensemble models achieve superior results compared to the state-of-the-art models while requiring less parameters and proposing a more hardware-friendly approach. On the Kvasir-Capsule dataset, an AUC score of 76.86% and a sensitivity of 60.65% is achieved. Compared to the baseline [Sá+23], 5 percentage points (pp) of more true positives were detected. Using the Galar dataset, the best AUC score of 76.98% was obtained with the ensemble model, including the SVM, while the highest sensitivity was reached with the ensemble model including the random forest model. However, considering the results of both datasets, the choice of the final classifier (SVM or random forest) does not seem to have a huge impact on the final performance. Since the Galar dataset was just recently published, no results targeting anomaly detection were present up to this date. In conclusion, the presented results can function as a baseline for future work.

The autoencoder in itself performed notably better with the Kvasir-Capsule dataset (AUC score: 64.70%) than the Galar dataset (AUC score: 45.83%), which might be explained by the large amount of unlabeled data in the Kvasir-Capsule dataset, that is missing in the Galar dataset. While the Galar datasets comprises millions of labeled images, it is still less compared to the numerous unlabeled images in the Kvasir-Capsule dataset. If the data at hand simply is not enough, it might prohibit the exploitation of the autoencoder's advantages to some degree, which mainly relies on training with a large quantity of unlabeled data. In contrast to that, the vast amount of annotated data in the Galar dataset might explain the strong classification results of the standard image classifier, which only relies on labeled data.
To evaluate the influence of the individual classifiers more closely, Figure 3.9 displays the ensemble model results involving the random forest classifier validated using the Kvasir-Capsule dataset, showing the classifier's logit distance plotted over the log(MSE) of the autoencoder.

The blue area indicates the region in which samples are classified as anomalies by the ensemble model. The substantial amount of True Positives (TP) is accompanied by a large fraction of False Positives (FP). Notably, when applying VCE, clinicians must inspect the classified samples in any case. Thus, this work focused on a high number of TP and a large sensitivity, while a certain amount of FP leading to a decreased specificity is accepted. It is further demonstrated that the model has learned to associate a large log(MSE) as produced by the autoencoder with an anomaly. Additionally, the Figure 3.9 indicates that the ensemble considers the classifier prediction strongly, with a slight bias towards

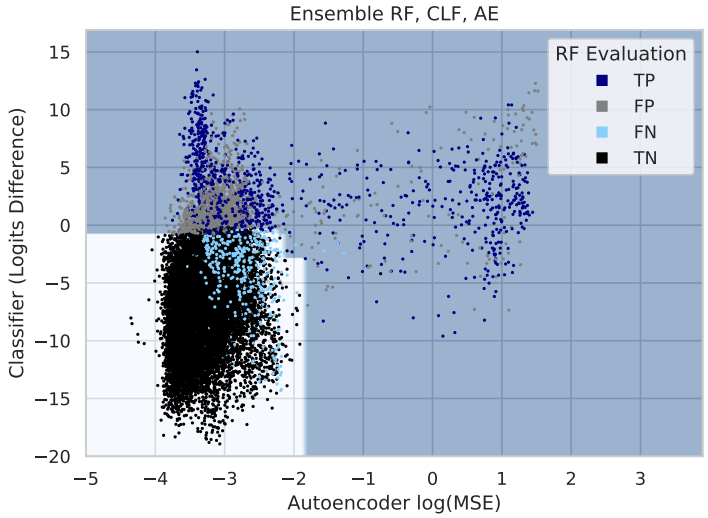


FIGURE 3.9: Ensemble model results including the random forest classifier, with the logit distance of the image classifier over the log(MSE) of the autoencoder. The evaluations are labeled by color as predicted by the ensemble model. The blue area displays, which samples were classified as anomalous, fromA3 [Wer+25a].

the detection of anomalies. Importantly, compared to current baselines, the model size was reduced from 60 million to maximally 5 million parameters through this approach. Hence, for both datasets, the designed ensemble model obtains superior results compared to current VCE baselines while depending on fewer parameters.

Additional Results

Besides the employed ensemble model from the published results, numerous other methods from the field of machine learning, specifically targeting anomaly detection, were explored. In the following, the most relevant approaches are briefly outlined and discussed.

One very promising method for successful anomaly detection is the deep Semi-Supervised Anomaly Detection (SAD) method from [Ruf+19], which is based on the idea that there is a lower entropy of distribution of normal samples in the latent space of an autoencoder compared to anomalies. In more detail, an autoencoder is first trained by minimizing the MSE, such that the neural network learns to encode and then reconstruct a given input image. Next, an embedding part of the neural network is trained together with the encoder, while the decoder is discarded. The first step in this retraining is to compute the mean, i.e. the center, of the embedded samples. Subsequently, the network is trained with the deep SAD loss $\mathcal{L}_{SAD}$ (see 3.1, as defined by [Ruf+19]), with n unlabeled samples $x_1, \ldots, x_n \in \mathcal{X}$ with $\mathcal{X} \subseteq \mathbb{R}^D$ in addition to m labeled samples $(\tilde{x}_1, \tilde{y}_1), \ldots, (\tilde{x}_m, \tilde{y}_m) \in \mathcal{X} \times \mathcal{Y}$:

$$\mathcal{L}_{SAD} = \min_{\mathcal{W}} \frac{1}{n+m} \sum_{i=1}^{n} \|\phi(x_i; \mathcal{W}) - c\|^2 + \frac{\eta}{n+m} \sum_{j=1}^{m} \left(\|\phi(\tilde{x}_j; \mathcal{W}) - c\|^2 \right)^{\tilde{y}_j} + \frac{\lambda}{2} \sum_{\ell=1}^{L} \left\| W^\ell \right\|_F^2 . \tag{3.1}$$

The form of the loss was directly taken from the original paper [Ruf+19]. Here, $\mathcal{W} = \{W^1, \ldots, W^L\}$ denotes the set of weights of the network with $L-1$ hidden layers, $\phi(\cdot, \mathcal{W})$ denotes the neural network realization corresponding to the parameters $\mathcal{W}$. The vector c and the weighting factors $\eta > 0$ and $\lambda > 0$ are hyperparameters. Finally, the matrix norm

in the final summand $\|\cdot\|_F$ denotes the Frobenius norm. It further holds that $\mathcal{Y} = \{-1, +1\}$, with anomalies denoted by $\tilde{y} = -1$ and normal samples denoted by $\tilde{y} = +1$. The theory is that the anomalies are expected to be pulled more slowly to the center due to the inverse in the loss function, such that preferably only the normal samples center around the previously defined mean. However, the AUC score never surpassed 60% on the Kvasir-Capsule dataset, and it was apparent, that the model was not able to clearly separate the normal samples from the anomalies in the latent space. Additionally, substituting the embedding part of this approach with different sizes of linear layers to modify the overall complexity was explored along with the addition of a SVM, aiming for another feature with non-linear kernels. However, modification of the embedding part did not substantially change the classification performance. This might be explained by the fundamentally different types of anomalies, which are detected by [Ruf+19] in comparison to the pathologies in the VCE datasets used in this work. For instance, objects from a randomly determined class of the CIFAR-10 [KH+09] or the MNIST [LeC98] dataset presumably differ much stronger from the rest of the dataset than a tiny anomaly in an image of otherwise healthy mucosa, as present in the VCE datasets. This potentially leads to a stronger separation of the anomaly class in the work of [Ruf+19] compared to data originating from the GI tract, in which the distinction is less clear. In the future, these experiments should be repeated with a combination of multiple VCE datasets. Nevertheless, it cannot be ruled out that this approach simply does not fit the given problem involving VCE image data.

To improve the supervised classification results with the Kvasir-Capsule dataset, the amount of annotated data was increased. This can be realized by pseudo-labeling, which was firstly introduced by [Lee+13] and aims to extend an existing dataset iteratively with predictions for unlabeled data. This approach was further refined by [Soh+20] with the FixMatch procedure. This adds various augmentations and a confidence value τ to the pseudo-label algorithm, in order to produce labels only if the confidence of the prediction succeeds the predefined τ and thus, theoretically prohibits a larger number of incorrect predictions. Furthermore, uncertainty-aware pseudo-labeling was introduced by [Riz+21], which adds an uncertainty factor κ to the confidence value. This considers the possibility that overall very confident, but ultimately incorrect predictions, might impair the overall performance. Hence, pseudo-labeling was applied to the unlabeled frames in the Kvasir-Capsule dataset to leverage the vast number of unlabeled images. First, a ResNet18 was trained with all existing labeled frames and subsequently evaluated on the unlabeled data. The predictions of the network were then used to label the unlabeled frames with pseudo-labels. Finally, the model was retrained with the labeled and pseudo-labeled data jointly, offering access to significantly more data. The uncertainty-aware pseudo-labeling results in a F1-score of 61.28%, which is marginally lower than the baseline with supervised training (62.57%). As a consequence, this was not further investigated, since the chosen mechanisms to prevent incorrect pseudo-label were not sufficient and did not enhance the overall classification performance.

3.2.3 Multi-task Model for VCE (Publication A4)

Publication A4

Title: Seeing More with Less: Video Capsule Endoscopy with Multi-Task Learning

Abstract: An ongoing challenge of the Video Capsule Endoscopy procedure remains the limited battery lifetime involving the risk of exhaustion before the entire small intestine is medically examined. Incorporating artificial intelligence into such capsule can provide smart real-time analysis and decisions, leading to a prolonged battery lifetime, an increased probability of covering the whole region of interest and the ability to perform decision on-site upon anomaly detection. This work presents a multi-task neural network that can jointly perform image-based organ classification for precise localization and basic anomaly detection that allows implementing additional features, while considering the resource constricted target devices. Validated on the Galar dataset, the approach outperforms current single-task models while limiting the total number of parameters to only 1 million. Overall, this presents a significant improvement towards an enhanced AI-based video capsule endoscopy.

Contribution of Authors

Julia Werner:	Conceptual design and implementation. Main author of this publication.
Oliver Bause:	Scientific advise and proofreading.
Maxime Le Floch:	Provision of the dataset.
Franz Brinkmann:	Provision of the dataset.
Jochen Hampe:	Secondary supervisor and provision of the dataset.
Oliver Bringmann:	Supervisor and scientific guidance.

State of the Art

While various multi-task methods have been presented targeting the GI tract in general, the small intestine is usually not the main objective in these studies as required for the VCE. For example, [Kon+21] focused on Crohn's disease detection with a multi-task approach using deep neural networks (ResNet50 and DenseNet121) on a private dataset involving 15 patients. Additionally, [Yu+21] used multi-task models to jointly learn segmentation and classification tasks to detect esophageal lesions. [Tan+23] performed polyp detection within the colon using a multi-task transformer model. More recently, [Xu+24] employed magnetically controlled capsule endoscopy to analyze gastric anatomical sites and lesions. Importantly, they discussed the model complexity and limited the overall model size by using a MobileNet with approximately 4 million parameters. However, these studies mainly focused on the esophagus, stomach or colon.

Approaches targeting specifically the small intestine using VCE images are predominantly limited to single-task models for either GI organ classification or anomaly detection. Given the inherent single-task focus of the previously published Rhode Island [Cha+22] (for anatomical regions) and the Kvasir-Capsule [Sme+21] dataset (for pathological sites), the multi-task options were limited. Nevertheless, since the Galar dataset [Le +25]

was published in 2025, which comprises anatomical as well as pathological labels, a database particularly well-suited for a VCE-based multi-task approach exists. This publicly available dataset allows conducting localization based on anatomical classification along with additional anomaly detection.

Approach

To prolong the "battery life" of a video capsule using AI-based decisions, a resource-constrained model considering the target architecture was provided, which is capable of precise localization within the GI tract and has basic anomaly detection features. Importantly, image-based organ classification can facilitate the localization of the capsule and the transmission of images should only be started when the small intestine is entered. This can save energy which would otherwise be needed to transmit images originating from the esophagus or stomach until the stomach exit is reached. Upon anomaly detection, additional features can be employed, such as an increased frame rate or resolution, targeting an enhanced analysis. Notably, when designing AI-based decision models for such tiny sensor edge devices, a basic hardware suitability of the model should be considered. A common multi-task learning method is hard parameter sharing that involves the sharing of hidden layers between multiple tasks while using task-specific output layers [Car93]. Leveraging this approach, a single, lightweight neural network, that can perform localization

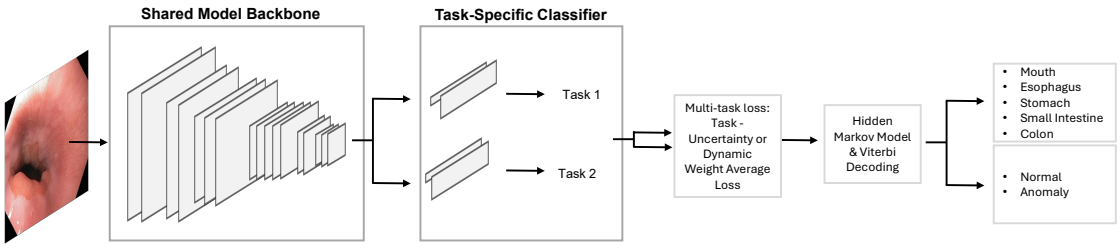


FIGURE 3.10: Multi-task approach involving a lightweight CNN with hard parameter sharing, task-specific classifier heads and post-processing with Viterbi decoding, from A4 [Wer+25b].

For both tasks, the architecture includes a shared backbone, which resembles the MobileNetV3-Small architecture. The model was then concluded by two task-specific classifier heads. To train this CNN, various established multi-task losses were tested, including the homoscedastic task uncertainty loss based on [KGC18], the Dynamic Weight Average (DWA) [LJD19] and the focal loss [Lin+17] to involve the class imbalance between normal and abnormal samples. The Galar dataset [Le +25] was adjusted, so that for each VCE input image, a label for the organ section as well as for an anomaly or healthy samples is present. Finally, for the organ classification, post-processing was performed with a HMM and Viterbi decoding to find the most likely sequence of organs as the capsule traverses the GI tract.

Results and Discussion

In the following, the multi-task results for both tasks are presented and compared to current baselines. The two main objectives were to improve the image classification performance

of the multi-task model compared to the single-task models while the second objective was to outperform current baselines for both tasks while additionally requiring less parameters than the baselines. The first objective was fulfilled, since the multi-task model generally performed better than the single-task models. One possible explanation for this is that the multi-task model better counteracts overfitting due to a regularization effect provided by the two learning objectives instead of one, which especially applies to the anomaly detection task. Nevertheless, this was also observed for the localization task, for which the performance was additionally boosted by the application of Viterbi decoding. Detailed results can be found in the Publication A4 [Wer+25b]. Table 3.13 shows the classification performance of the presented multi-task results in comparison to the current baselines for both tasks.

Table 3.13: Results of the multi-task model compared to current baselines.

Task	**Approach**	**Accuracy** [%]	**F1-Score** [%]	**Params**	**MAC**
Localization	Single Task Baseline [Le +25]	81.0	71.0	25 M	4 B
Localization	Multi Task CE & HMM	92.99	86.52	1 M	64 M
Localization	Multi Task UW & HMM	90.24	88.84	1 M	64 M
Anomaly Detection	Single Task Baseline [Wer+25a]	87.28	37.01	4 M	189 M
Anomaly Detection	Multi Task CE	84.05	53.08	1 M	64 M
Anomaly Detection	Multi Task UW	82.09	52.19	1 M	64 M

Compared to [Le +25], the accuracy was improved with both multi-task approaches, while reducing the required number of parameters and MAC operations. Employing the task uncertainty weighted loss (UW) with the multi-task model leads to an accuracy of 90.24% and the highest F1-score of 88.84% in comparison to 71% of the baseline. The F1-score of the multi-task model with the cross-entropy (CE) loss is slightly lower than for UW, with 86.52%. However, this approach still obtains a higher accuracy of 92.99% compared to the baseline with an accuracy of 81%. It can be concluded that the multi-task models outperform the current baseline with either employed loss functions. Additionally, only 64 M MAC operations were needed in contrast to 4 B MAC operations required for the baseline model. Thus, the enhanced classification performance along with the reduced number of MAC operations yields a better suited model for the VCE application than current baselines.

In addition, the presented model is not only superior regarding the localization task, but can also perform anomaly detection within the same single model. However, the low F1-scores of 52.19% and 53.08% indicate that the model misclassifies various data samples and also exhibits some randomness. This could result from a varying image quality regarding the camera angle and lighting, potential misclassification or air bubbles as well as digestion remains impairing the visibility. While these factors apply to the organ detection task as well, they can have a stronger impact on the anomaly detection task due to less annotated data and a higher labeling variability depending on the annotator. Nevertheless, since the localization performance is not impaired but instead improved with the proposed approach compared to current baselines, the anomaly detection capabilities are an additional benefit while limiting the total model size compared to the literature baselines.

Additional Results

In the following, further insights into the different results of the CNN on its own in comparison to the hybrid approach with the HMM and Viterbi decoding are shown by inspecting the classification performance on an individual patient level. Furthermore, new results combining parts of the anomaly detection approach presented in Section 3.2.2 and

the presented multi-task model in this Section are introduced. Figure 3.11 displays a comparison of CNN predictions and the HMM and Viterbi approach for each existent patient in the Galar test set.

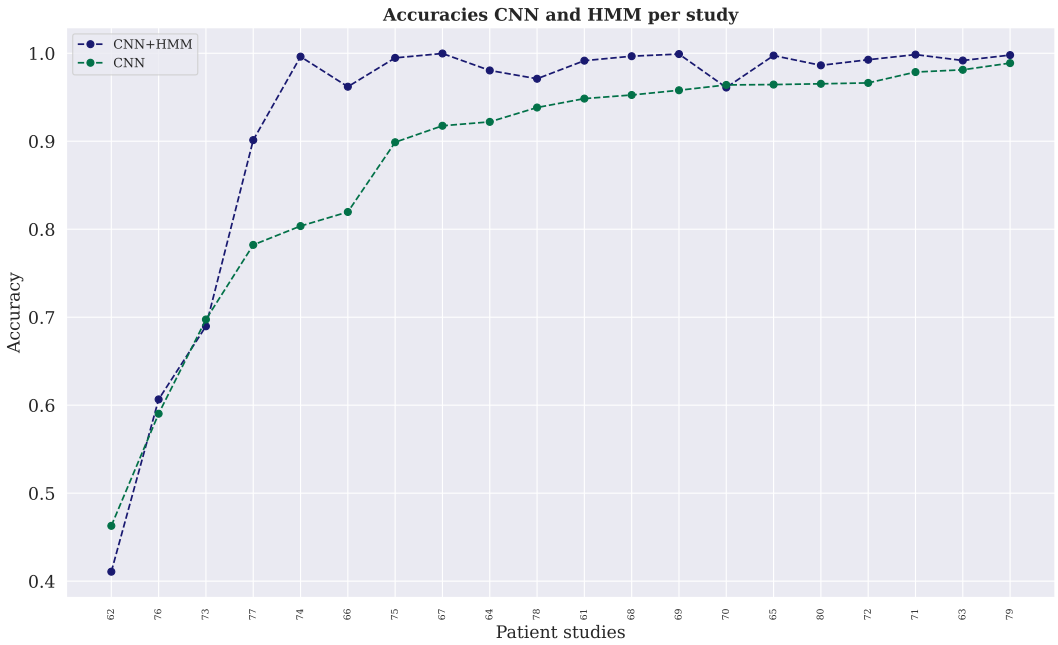
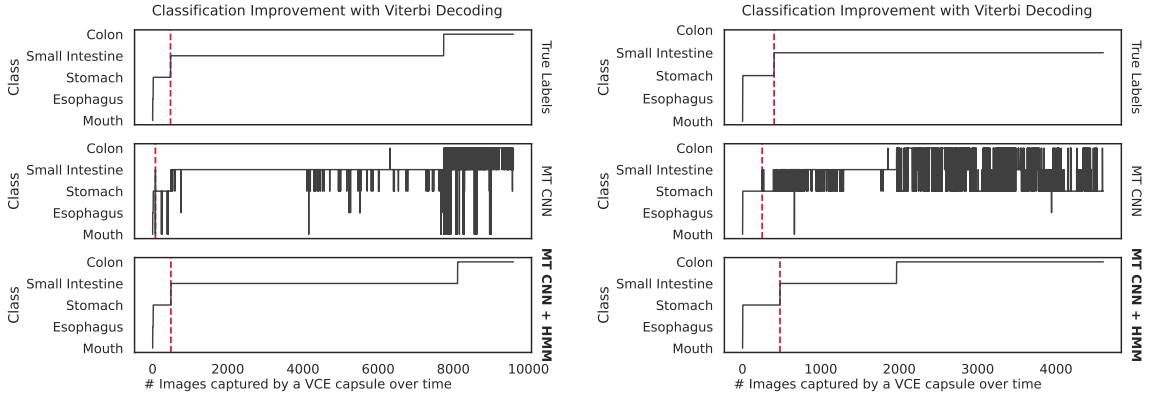


FIGURE 3.11: Prediction comparison of the CNN compared to the hybrid approach of CNN, HMM and Viterbi decoding validated using the Galar dataset with the multi-task model and the DWA focal loss function.

This demonstrates the superior performance of the hybrid approach of Viterbi decoding and a neural network for this time-dependent classification problem. The higher the accuracy of the CNN, the better does the HMM with Viterbi decoding perform. However, there are also some outliers present. More specifically, for patient 62 neither the CNN nor the hybrid approach performed well, with an accuracy of approximately only 40%. Thus, the evaluation for this patient is examined more closely in the following.

Figure 3.12 displays for the first two patients of the Galar test set the classification performance of the multi-task CNN itself in comparison to the multi-task model in combination with the HMM. Hereby patient 61 functions as an example for whom the approach effectively classifies the data and patient 62 provides an example for whom some challenges remain.

For both patients, the plots emphasize the structural improvement realized by the HMM and subsequent Viterbi decoding. Even though, the overall accuracy declines for patient 61, using subsequent Viterbi decoding is beneficial for this application due to its filtering effect and structural improvement. The red line in the plots indicates the first detection of the small intestine. For the two examples, the CNN alone detects the small intestine too early, exhibits many misclassifications and jumps between the different organs. In contrast, the classification performance of the HMM and Viterbi decoding is characterized by a clean transition between the different organs without allowing a switch to a previous organ. However, in cases in which the CNN already performs inferior as shown in 3.12b for patient 62, the Viterbi decoding still leads to smoothing of the predictions, but might predict a transition to the next organ too early (in this case the transitioning from small intestine to colon), if the CNN has low confidence on a definitive class.



(A) VCE study ID 61.

(B) VCE study ID 62.

FIGURE 3.12: Classification performance of two single VCE studies shown for the multi-task CNN compared to the multi-task CNN in combination with the HMM using the DWA focal loss function.

The discussed multi-task approach targeting anatomical classification as well as anomaly detection performs very well by employing a purely supervised approach using the MobileNetv3 architecture. Building upon the anomaly detection work which includes unsupervised machine learning techniques, it is aimed to enhance the sensitivity of detecting pathologies. To refine this method further, a natural next step is to combine the results from the anomaly detection manuscript [Wer+25a] from Section 3.2.2 and the multi-task publication [Wer+25b] presented in Section 3.2.3. Based on the discussed ensemble model, an autoencoder and random forest classifier were extended to learn the anatomical classes in addition to the anomaly detection to embed the multi-task possibility. The reconstruction error of the autoencoder was refined by incorporating three different losses, which consider the class imbalances: The Structural Similarity Index Measure (SSIM) [Wan+04], the Part-Based Structural Similarity Index Measure (PSSIM) [Shi+22] and the multiscale SSIM [WSB03].
The SSIM is computed between two pixel values of an image and then averaged over the whole image to assess the structural similarity, while the PSSIM extends the SSIM loss by taking the unique individual patterns of a person into account while considering only small patches of images. The reasoning behind this is that anomalies rather affect small parts of the images and by using only small patches, those parts are weighted stronger than if the whole image is included. The multiscale SSIM assesses the structural similarity depending on different viewing conditions, such as the display resolution and viewing distance. The autoencoder is then trained based on all three loss functions combined. Using a XGB Boosting classifier [CG16], which receives the output of the different losses as features, as a final evaluation method, a final accuracy of 83.75%, an F1-score of 60%, a precision of 58.32% and a recall of 70.18% on the Galar dataset and the anomaly detection task was achieved. Additionally, this model was capable of classifying the anatomical organs for the localization task with an accuracy of 92.36%, a F1-score of 89.63%, a precision of 86.95% and a sensitivity of 93.60% in combination with Viterbi decoding. These results indicate a significant improvement for the anomaly detection task with an increased F1-score of 7.81 pp, which might be caused by the refinement of loss functions. The performance on the localization task also improved, besides the difficulty of balancing those tasks. A possible next solution strategy is to formulate a combined loss that weighs both tasks equally and also considers class imbalances for both tasks. This must be cross-validated on unseen VCE data.

3.2.4 Mislabel Detection for VCE data (Publication A5)

Publication A5

Title: Reliable Mislabel Detection for Video Capsule Endoscopy Data

Abstract: Successful image classification by machine learning models strongly depends on the access to large and accurately annotated datasets. Specifically for medical vision datasets as obtained by the Video Capsule Endoscopy (VCE), such annotation is tedious and must be executed by experienced gastroenterologists. The labeling process is further complicated by ambiguous class boundaries, which complicate definite classifications. In this work, a framework to detect mislabeled data is introduced and validated on the Kvasir-Capsule and the Galar dataset to assess the image labeling quality of the available VCE data and clean the datasets from noisy labels. Additionally, the identified samples were reviewed by three specialized physicians to evaluate the performance of the introduced framework. This demonstrates that the presented framework accurately detects and filters noisy data. Furthermore, anomaly detection performed on the cleaned datasets outperforms current baselines that involve uncleaned VCE datasets.

Contribution of Authors

Julia Werner:	Conceptual design and implementation. Main author of this publication.
Julius Oexle	Implementation and evaluation.
Oliver Bause	Scientific advise and proofreading.
Maxime Le Floch:	Re-annotation of identified frames.
Franz Brinkmann:	Re-annotation of identified frames.
Hannah Tolle:	Re-annotation of identified frames.
Jochen Hampe:	Secondary supervisor and scientific medical guidance.
Oliver Bringmann:	Supervisor and scientific guidance.

State of the Art

In the past, there have been several endeavors to detect noisy (false) labels in datasets to improve the classification performance of neural networks [Kan+18; Xia+15]. For instance, [FQD20] and [Ost+18] used ensemble methods to clean datasets from erroneous labels. Other strategies focused on enhancing the robustness of machine learning models towards noisy data instead of correcting the labels themselves. Most prominently, this is performed by modifying the loss functions to be less sensitive to erroneous gradients caused by noisy samples [Pat+17; Wan+19; ZS18]. Other approaches rely on the assumption that similar samples also resemble each other in the feature space and should therefore have the same annotation [Lee+18; Sha+20]. Accordingly, deviating samples within the feature space that have the same label are assumed to be noisy.

For neural network training, the presence of large datasets is crucial. However, particularly in the field of medicine, annotations can often only be provided by experienced physicians, such that large amounts of annotated data remains sparse. Additionally, VCE datasets

are subject to incorporating a huge class imbalance with the pathologies being in the minority [Le +25; Sme+21]. If further noisy labels among the anomalies exist in such imbalanced datasets, this can have a particular impeding influence on the classification performance using deep learning models. To tackle the anomaly detection based on VCE data, in 2025, the Galar dataset [Le +25] has been published, comprising the largest number of annotated multi-label VCE images. Although this dataset is very large with 3.5 M images, anomaly detection remains challenging due to the present class imbalance. For instance, in the original dataset publication, a F1-score of only 5% for polyp, 2% for erythema, and 14% for blood detection was achieved [Le +25]. This was further enhanced to a F1-score of 37% and 54% for a joint pathology detection in [Wer+25a] and [Wer+25b], respectively. While this presents a substantial improvement, it also indicates that the models still not manage to detect all pathologies. The objective of this work is to address this from a different perspective. The main strategy is to identify potential mislabels within the available VCE datasets and clean the datasets from such noisy samples instead of primarily designing robust neural networks and loss functions. To demonstrate the effectiveness of this approach, using the cleaned dataset, anomaly detection is performed employing a standard neural network and the results compared to current baselines.

Approach

The successful identification of erroneous labels within the available VCE datasets is complicated by a missing ground truth. It is not known, how many noisy data samples exist in the available datasets. To circumvent this problem, this work follows a two-stage design: first a controlled experiment using the Kvasir-Capsule dataset, secondly the application of the pipeline to the Galar dataset, followed by a thorough review. For the Kvasir-Capsule dataset, we make the assumption that noisy labels are very unlikely for two reasons. First, as only a selected and comparably small number of samples were annotated, it is assumed that this was performed very carefully. Secondly and more importantly, each pathological frame is accompanied by a bounding box, clearly indicating the position of the anomaly. It is assumed that this increases the likelihood of an anomaly being actually on the image. For simplification, it is assumed, that the datasets consists only of correctly labeled data. Thus, for the Kvasir-Capsule dataset labels are randomly flipped to introduce noisy labels manually. After the detection framework was applied, the effectiveness of mislabel detection is evaluated. On the other hand, for the Galar dataset, the main objective is to identify pre-existing erroneous annotations. Therefore, 100 samples that were identified as noisy by our pipeline were subsequently reviewed by the Co-authors 4, 5, and 6 of this manuscript, who are experienced gastroenterologists. These 100 samples serve as a small sample size and a first indication for the reliability of the proposed framework.

The framework for mislabel detection is thereby centered around a Gaussian Mixture Model (GMM) [MP00; PFJ06] since they have been successfully used for the detection of noisy samples [LSH20]. This probabilistic model tends to obtain an increased loss for mislabeled data points compared to correctly labeled data, which is exploited in our pipeline. As visualized in Figure 3.13, the cleaning process starts by training a neural network on the uncleaned dataset with subsequent GMM training. To determine the probability of a label being correct, the neural network's confidence and prediction are used. Based on the highest noise reduction, the first k^c labels are corrected. This is followed by three CNN and GMM training and a filtering step, in which the first k^f labels that obtained the highest noise probability are filtered and cleaned. Based on the work of [Jia+24], the correction step was fused with the filtering step and for each sample decided whether to include it in the cleaned dataset with a corrected label or to filter it out.

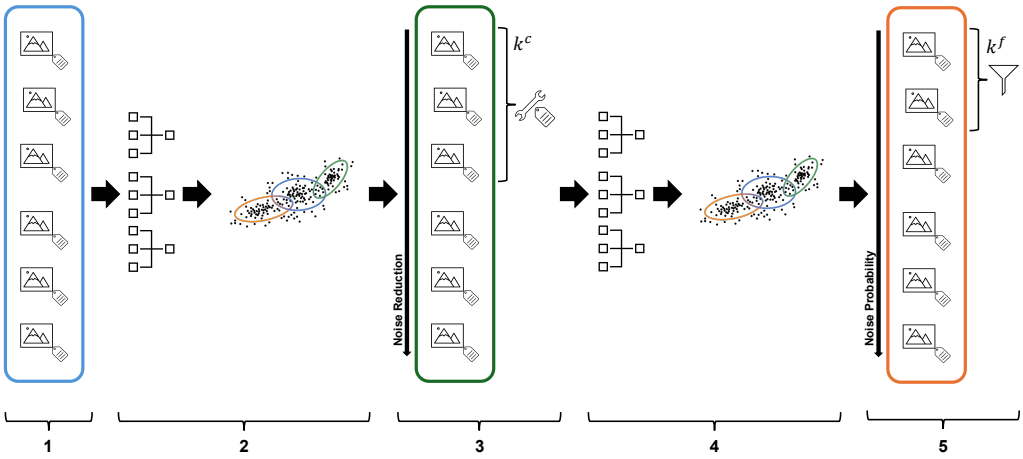


FIGURE 3.13: Pipeline to identify, correct and filter mislabeled data: starting with an uncleaned dataset (1), three CNN trainings with subsequent GMM training (2), a correction step (3), training to determine the noise probability (4), and finally a filtering step (5) (from A5).

To consider the required low model complexity, when targeting an application such as the VCE which involves resource-limited embedded devices, a MobileNetV3 [How+17] is selected and trained on the cleaned dataset to perform anomaly detection. Furthermore, a focal loss [Lin+17] is used to address the present class imbalance. Finally, the results are compared to current baselines to assess whether a preceding dataset cleaning improves the detection of pathologies in the GI tract.

Results and Discussion

In the following, the results obtained by the mislabel detection pipeline are summarized, starting with the Kvasir-Capsule and followed by the Galar dataset. Finally, the anomaly detection results using a cleaned dataset are discussed and compared to current baselines.

For the Kvasir-Capsule dataset, the majority of intentionally induced noisy samples were correctly detected by the introduced framework with an accuracy of $\approx 96\%$ for 5% injected noise. In absolute terms this corresponds to 2262 detected noisy labels of 2360 samples in total. Furthermore, a t-Distributed Stochastic Neighbor Embedding (t-SNE) plot (not shown) visualized that the implemented pipeline primarily corrected labels if their latent representation resides in the cluster of the other class or within a transitional region between those two classes. Accordingly, after the cleaning pipeline was applied, more coherent clusters were observable. In total, $199,359$ samples from the Galar dataset were identified as noisy samples, from which $167,709$ samples were filtered (4.8% of the whole dataset) and $31,650$ (0.9%) labels corrected. In absolute terms, these are very large numbers, that might explain the preceding difficulties when performing anomaly detection.

To provide further clarification in this matter, a MobileNetV3 was trained on the cleaned dataset and the main anomaly detection results shown in Table 3.14.

This demonstrates that training on a cleaned dataset indeed significantly boosts the classification results, with a F1-score of 71.58% compared to using the uncleaned dataset (F1-score of 53.70%). Notably, the benefit of training on a cleaned dataset is further

Table 3.14: Anomaly detection results on the Galar dataset compared to current baselines.

Cleaned dev set	**Accuracy** [%]	**F1-Score** [%]
Uncleaned [Wer+25a]	87.28	37.01
Uncleaned [Wer+25b]	87.7	54.38
Uncleaned	90.99	53.70
Filtered	93.83	71.58

reinforced as this approach outperforms both baselines [Wer+25a; Wer+25b]. This becomes even more meaningful when considering that the employed model is not nearly as complex as the approach provided by [Wer+25a]. Applying an ensemble model to the cleaned dataset and thus, combining both strategies, might further enhance the current results. Additionally, 100 selected samples, that were identified as noisy by our pipeline were re-annotated by experienced clinicians. They confirmed that 78% of the identified labels were indeed incorrectly labeled, which strengthens the effectiveness of our approach.

To conclude, an effective mislabel detection pipeline was designed and implemented. The results suggest that applying a cleaning pipeline to VCE image datasets is beneficial before conducting image classification. While it is advisable, to re-annotate all identified noisy samples by gastroenterologists, the provided corrected annotations and filtered dataset splits as produced by this framework on its own yield a substantial enhancement for the anomaly detection within the GI tract.

3.2.5 Raw Image-Based Localization and Hardware Simulation (Publication A6)

Publication A6

Title: Smart Video Capsule Endoscopy: Raw Image-Based Localization for Enhanced GI Tract Investigation

Abstract: Deep neural networks employed for image classification can quickly become too large and complex, involving various operations, which usually exceed the capabilities of low-power sensor edge devices used in real-world applications. Additionally, images are commonly captured in Bayer pattern format by miniature cameras and are subsequently converted to the standard Red Green Blue (RGB) format, which is also typically used for neural network training in the field of machine learning. If the energy for this conversion step was saved, for resource-constrained devices, this spared energy could be used for other important tasks. By means of the Video Capsule Endoscopy, the presented work targets the provision of an AI-based hardware-suitable low-resolution image classification approach involving a hardware-aware, quantized CNN directly trained on raw Bayer images in combination with quantized Viterbi decoding. The proposed approach yields an accuracy of 93.06% using only 63,000 parameters. The method is additionally transferred to a customized PULPissimo System-on-Chip with a Reduced Instruction Set Computer (RISC)-V core, a miniature NanEyeC camera sensor and the low-power hardware accelerator UltraTrail to demonstrate the energy-efficient AI-based image classification pipeline. With this prototype, this work presents extensive results on the energy demands of such VCE system involving a machine learning classifier.

Contribution of Authors

Oliver Bause:	Conceptual design and implementation of the hardware demonstrator. Shared first authorship.
Julia Werner:	Conceptual design and implementation of the quantized CNN, HMM and Viterbi evaluation in software. Shared first authorship.
Paul Palomero Bernardo:	Review and scientific advise.
Oliver Bringmann:	Supervisor and scientific guidance.

State of the Art

Small, sensor edge devices involving a camera generally capture images using a Bayer color filter. Subsequent demosaicing and denoising can be for instance performed by using neural networks [Kha+21; KPK23]. However, whether the usage of CNNs is applicable on low-power mobile devices for this task is not guaranteed due to their typically high computational demands. Notably, evaluating CNNs directly on raw Bayer images can not only circumvent this problem, but also potentially presents a promising approach to yield equivalent results involving less computation steps by omitting prior conversion to RGB images. However, only very little research has been conducted in this matter. For example, [CL21] applied a SqueezeNet directly to Bayer images for the classification of

hand postures. When using an additional demosaicing technique, they achieved proficient results of 98.28% accuracy. Nevertheless, without this technique and with direct usage of raw Bayer images, the accuracy dropped to less than 25%. This strongly indicates that there is room for improvement and poses the question if CNNs are able to recognize the underlying structure and characteristics of given data purely based on raw images in Bayer pattern format.

Substantial progress has been made on VCE image-based organ classification with accuracies between 69.84% [Abi+25] and 97% [Cha+22]. [Wer+23] additionally yielded a strong classification performance with 96.95% accuracy with a parameter reduction from 56 M to 1 M. In 2025, [Pal+25] presented first hardware simulation results with a simple CNN, which is executed on a low-power programmable edge AI accelerator. Importantly, all research was conducted using RGB images, requiring an additional image conversion step if actually applied in a VCE study. Onboard real-time VCE image evaluation was firstly explored by [Sah+22], using a model with 3 M parameters. In spite of that, they report an operation time of maximally an hour (h) due to a high power consumption, which is significantly less than the 8 to 12 h of current VCE devices. Instead of shortening the battery lifetime, adding AI to such capsules is expected to prolong the operation time in this work.

To conclude, the main objective of this publication is to evaluate VCE images directly using raw Bayer images to save the conversion step, provide extensive results with a hardware demonstrator on the energy consumption of this system and to evaluate how this impacts the overall battery lifetime.

Approach

Directly processing raw Bayer images is expected to save electrical energy and chip area which would be otherwise needed for the conversion to RGB images before conducting classification with a neural network. As Bayer images only contain one color channel per

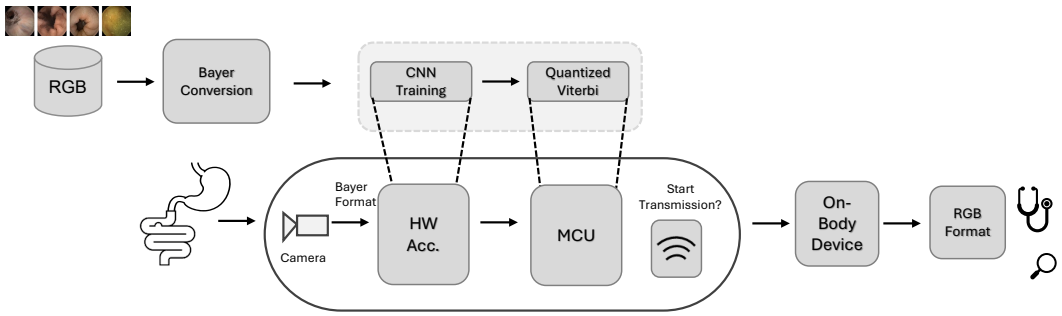


Figure 3.14: Processing of raw Bayer images with a quantized CNN and Viterbi decoding on a PULPissimo-based demonstrator with a RISC-V core, a NanEyeC miniature camera sensor, the UltraTrail AI accelerator and a Microcontroller Unit (MCU) for hardware simulation (from A6 [Bau+25]).

This image classification pipeline is combined with a VCE demonstrator, involving a NanEyeC camera sensor, the UltraTrail hardware accelerator for CNN execution, and a Microcontroller Unit (MCU) on which the quantized Viterbi algorithm is conducted. This functions as a decision pipeline, which can determine a reasonable start of transmitting VCE images and additionally adapt the frame rate depending on the current localization. The described system should assist in prolonging the battery lifetime of VCE devices and ensure that the battery is not depleted before the entire small intestine has been covered. For training, validation, and testing, the Rhode Island dataset [Cha+22] was used along with the MobileNetV3-small architecture to evaluate how an established network handles the different image types. Finally, a lightweight feed-forward CNN with only 67,000 parameters was used, which can be directly employed on the demonstrator as previously demonstrated by [Pal+25] for this VCE setting. Following [Wer+23], a HMM and Viterbi decoding was used for final post-processing. This was additionally extended to a purely quantized computation in fixed-point representation, ensuring its hardware suitability.

The used system architecture involves a customized PULPissimo System-on-Chip (SoC) featuring a RISC-V core. It was synthesized in 22FDX+ technology, and equipped with the UltraTrail AI accelerator [Ber+20], which enables real-time inferences of convolutional neural networks. The SoC further consists of an integrated hardware controller, that ensures the communication with an asm NanEyeC miniature camera sensor [AG24]. To simulate a VCE system, the employed camera has a Red Green Green Blue (RGGB) Color Filter Array (CFA) applied, exhibits a size of $1\,\mathrm{mm}^2$ and captures images with a resolution of 320×320. Overall, this architecture should function as a basic prototype including the most fundamental elements of a VCE system.

Results and Discussion

The presented work explored whether usage of raw images compared to the standard RGB approach impairs the image classification results. Table 3.15 presents results of a MobileNetV3 trained and tested either with images in Bayer pattern format or with RGB images compared to current literature results for the Rhode Island VCE dataset [Cha+22].

Table 3.15: Comparison of the MobileNetV3 trained and tested on raw Bayer Pattern images (RGGB) vs. RGB images compared to the literature.

	Input	**Accuracy [%]**	**F1-Score [%]**	**Params**
Inception ResNetV2 [Cha+22]	RGB	97.10	97.13	56 M
Swin Transformer [Abi+25]	RGB	69.84	69.85	195 M
MobileNetV3 [Wer+23]	RGB	96.95	-	1 M
MobileNetV3	RGB	97.14	90.89	1 M
MobileNetV3	RGGB	96.20	88.12	1 M

Compared to the literature results, the MobileNetV3 performs as well as the larger models while needing only less than 2% of the number of weights compared to the Inception ResNetV2 and approximately 0.5% of the number of trainable parameters used in a Swin transformer. A possible explanation for this observation is that the smaller model might be less prone to overfitting, which then might lead to a comparably strong performance considering the model size reduction. Furthermore, the accuracy is only slightly reduced from 97.14% to 96.20%, if images in Bayer pattern format are used. This classification

impairment might be caused by different noise properties in the different images. When a raw Bayer image is converted to a RGB image, noise within the image is altered depending on the demosaicing and denoising method. However, this change in information by RGB conversion might also produce worse results in some cases as shown in the following (Figure 3.15).

To further reduce the model size, the experiments were repeated with an even less complex feed-forward model suitable for low-power hardware architectures based on [Pal+25], consisting of only 62,976 parameters. The model was trained with quantization-aware training and the classification performance with different word widths ranging from 2 to 10 bit per weight explored for standard RGB compared to Bayer pattern images (RGGB) in Figure 3.15.

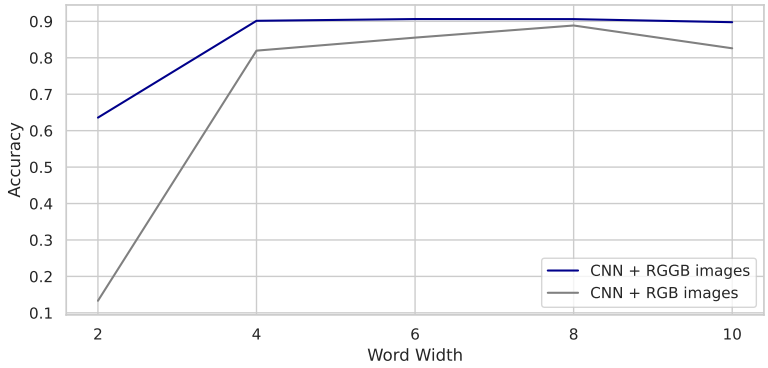


FIGURE 3.15: Accuracy of the hardware-suitable CNN tested on RGB and RGGB using different word widths (from A6 [Bau+25]).

In contrast to the MobileNet experiments, this demonstrated for one, that the performance was not inferior if trained directly with RGGB images. Thus, it cannot be generally concluded that directly using raw Bayer images as an input leads to worse results as might be the first impression when inspecting the MobileNetV3 results. Extensive comparison of various image datasets using a wide range of neural networks is required to make a general statement in this matter. However, in this work, the objective was to discuss the effects of skipping a conversion step prior to CNN evaluation only for the VCE setting. This indicates that processing the data directly as captured by camera sensors, does not impair the results and the network is still capable of learning the relevant features. Only a word width smaller than 2 bit led to significantly worse results. Hence, using only 2 bit might simply be too restrictive for this task. Notably, a word width of 8 bit provides a good trade-off between accuracy and complexity and was employed in the following. Post-processing with Viterbi decoding resulted in an additional classification enhancement with an accuracy of 90.61% compared to 93.06% and a more accurate stomach exit detection.

To assess hardware suitability and evaluate whether the approach is indeed eligible for a VCE system, hardware simulation needs to be conducted and the electrical energy consumption for each component evaluated. Using the PULPissimo System-on-Chip demonstrator, the image processing was simulated and the input power and electrical energy consumption results along with the inference time for each component shown in Table 3.16.

The consumed electrical energy for processing images with Viterbi decoding is with 0.17 µJ comparably low when set against the energy demand of the image sensor, the LEDs or the MCU and the accelerator. This was expected due to the low number of required operations

Table 3.16: Energy and power consumption of the presented demonstrator shown for each module per frame.

	Image Capture			CNN	Viterbi
	Image Sensor	**LEDs**	**MCU**	**MCU & Acc.**	**MCU**
Input Power [mW]	8.51	14.78	7.23	16.63	9.94
Inference Time [s]	12.79	12.79	12.79	0.31	0.02
Electrical Energy Demand [µJ]	108.93	189.15	92.56	5.14	0.17

for analyzing a frame during Viterbi decoding compared to the number of operations performed by a CNN. Furthermore, if the frame rate is decreased accordingly, this work showed that the power consumption can be reduced by almost 90% with this approach compared to a setting without an AI-based decision pipeline. As shown, this approach is capable of real-time organ classification on edge devices and therefore applicable for low-power on-site localization in VCE devices.

3.3 Additional Results Video Capsule Endoscopy

The preceding sections presented methods and results for image-based anomaly detection targeting the VCE along with precise organ detection and hardware simulation using a model directly on raw Bayer pattern images. Consequentially, a next step is to include additional information within the decision-making pipeline of the AI-based VCE setup to further improve its performance. Ancillary data could be gathered by various sensors, for instance, providing information on the capsule's speed, acceleration or position. However, considering the lack of such datasets, this would require careful collection of the respective data, which is not a goal of this thesis. When targeting the improvement of the VCE, this work focused on image-retrieved information due to the present image datasets. In addition to the existent and previously discussed annotations, the Galar dataset furthermore comprises annotations considering the view of the capsule and the overall image quality. Notably, this information only exists for six patients out of the 80 VCE studies. Therefore, in the following, using these respective six studies, experiments targeting the visibility assessment for the VCE are discussed, which complements current research offering additional insights regarding advisable interventions during such procedure.

3.3.0.1 Visibility Assessment for VCE

In addition to organ and anomaly detection within the GI tract, assessing the visibility of the capsule's captured images can be another important factor, when conducting VCEs. Obtaining information on the image quality can assist in deciding whether data should be transmitted (e.g. only high quality images) or whether it is reasonable to evaluate an anomaly detection model at a specific moment. If the visibility is rather worse, conducting anomaly detection probably yields inoperable results. In this case, it might be advisable to save the energy that is needed for model execution. Besides information on anomalies and organ sections, the Galar dataset [Le +25] contains annotated data indicating whether the capsule's view is *good*, *reduced* or *bad*. This partitioning was performed based on the visible fraction of an image, which can be altered by air bubbles, digestion remains or a general restricted field of view. Since up to this date, to the best of our knowledge other VCE datasets did not contain annotations for this matter, the Galar dataset is the first publicly available dataset, which enables view assessment based on VCE images. However, labeling for this specific visibility task is very time-consuming, which resulted in annotations for

only 6 VCE patient studies. Since such a small sample size is especially prone to overfitting, the influence of different data augmentation techniques on the classification performance was investigated to counteract potential overfitting in the following. Table 3.17 shows the results of the visibility task on the Galar dataset for different standard augmentations in comparison to a baseline without any augmentations and the original results by [Le +25] to evaluate if this augmentation enhances the overall model's performance.

Table 3.17: Impact of different augmentations on the visibility task (accuracy [%], precision [%] and F1-score [%] are shown).

	Visibility/View			
Augmentation	Accuracy	Precision	Recall	F1-Score
Baseline [Le +25]	54.0	–	–	54.0
Baseline w/o augmentation	86.83	63.21	63.37	63.27
RandomBoxBlur	85.38	76.29	51.25	56.41
RandomBrightness	73.06	64.52	59.31	59.32
RandomContrast	61.73	63.96	61.72	55.40
RandomGaussianNoise	85.80	74.53	50.12	55.72
RandomMotionBlur	87.08	74.67	57.85	63.43
RandomPerspective	83.88	66.46	49.89	52.66
RandomPosterize	86.41	75.58	52.17	53.71
RandomRotation	87.39	85.86	50.83	53.94
RandomSaltAndPepperNoise	84.55	78.25	58.42	58.36
RandomSaturation	88.85	84.06	58.69	60.67
RandomSharpness	84.76	73.04	57.08	60.88
RandomShear	86.22	73.98	52.71	58.42
RandomVerticalFlip	81.25	73.34	62.57	61.94
RandomHorizontalFlip	84.91	70.09	46.34	57.36
Combination 1 RandomMotionBlur, RandomRotation	**88.59**	70.47	66.26	**67.93**
Combination 2 RandomSaturation, RandomRotation	79.80	56.42	**69.62**	60.55
Combination 3 RandomMotionBlur, RandomSaturation	64.79	54.71	63.94	53.68
Combination 4 RandomMotionBlur, RandomSaturation, RandomRotation	86.78	**70.85**	63.31	65.52

It can be observed that the employed model size is significantly reduced compared to [Le +25], needing only 1 M parameters with the MobileNetV3 in contrast to 25 M parameters of the ResNet50, while outperforming the more complex model (F1-score 63.27% vs. 54.0%). This matches the previous observations obtained from the other classification tasks due to a lower susceptibility to overfitting of less complex models. For each augmentation type, a hyperparameter search over 30 runs was conducted to find the best parameter for each augmentation and to prevent differences merely based on more favorably hyperparameter choices for specific augmentations. Furthermore, custom augmentations were implemented, which added synthetic bubbles to the VCE images. However, this did not lead to any substantial improvement and was thus not further investigated. In general, most augmentations did only boost a single metric without substantially enhancing the overall prediction performance. One of the few augmentations that led to a general improvement was the combination of random motion blur and random rotation, resulting in an increment of the accuracy with 1.76 pp and of the F1-score with 4.66 pp.

Overall, the sparse number of only six patients was very prone to overfitting, and it was observable that simply adding augmentations did not structurally improve the training process. However, given the extremely tedious labeling process, anticipating a substantial

increase of the data amount is not feasible. Alternatively, upsampling techniques could be explored to tackle this problem further. Additionally, the annotations of VCE images differ slightly depending on the annotator. Whether the view is slightly reduced or still proficient and if there is no view at all or merely a strongly reduced view is perceived differently by multiple people. This labeling uncertainty is inherently embedded in the given annotations and might be adapted by neural networks during training leading to a diminished image classification performance which might further explain the impaired classification results.

3.4 Concluding Discussion

The obtained results demonstrate that embedded medical sensor edge devices can benefit from AI-based classification models. This thesis showed that machine learning models, if designed efficiently with the target hardware in mind can be well-suited for low-power architectures and enhance real-world medical systems. By employing a classification model and enable smart decision-making, the overall energy demand of such devices can be reduced, prolonging the battery lifetime or enabling the addition of new functionalities. For example, this can be realized by ensuring that neuroimplant simulation only occurs when seizures are accurately detected by the presented techniques, instead of stimulating constantly. In addition, transmitting only images of interest for the VCE based on the anomaly detection and localization models reduces the overall energy demand. Importantly, models with sequential dependencies, such as Recurrent Neural Networks (RNN) [RHW85] are accompanied by various obstacles with regard to the hardware design, parallelization and memory accesses. For example, as the computation in RNNs at each time step depends on the computation at the time step before, the parallelization of computations at multiple time steps of a single sequence is not possible. Furthermore, this is accompanied by an increased number of memory accesses to read and update their hidden states at each time step. The usage of input sequences of variable length additionally leads to an inconsistent memory usage.

Therefore, this thesis focused mainly on feed-forward neural networks when targeting ultra-low-power hardware accelerators. The presented and discussed approaches are useful methods for efficient seizure detection and VCE image classification and can be further adjusted for other problems involving 1D time-dependent data or medical images.

For both applications, developing such techniques was challenged by multiple factors. One very important aspect is potential mislabels. When only small number of labels are available, false annotations weigh more heavily. The labeling of medical data is not only tedious but also very challenging, even for trained physicians. Furthermore, in medical real-world datasets, some degree of label noise cannot be avoided due to variation between annotators and systematic errors [Kar+20]. Following this, the same pathology might be labeled differently by multiple clinicians. In respect to the VCE, the distinction between pathologies is not clearly outlined for all images; in addition, different physicians would probably determine the starting point of a seizure differently than others. This uncertainty of the annotators is presumably represented in the networks and lowers the overall classification performance. This sparse amount of labeled medical datasets of high quality has been determined to generally limit the supervised learning approaches for such medical vision tasks [Hol+20; Est+19; Lut+19]. In this work, this was counteracted by designing a framework capable of the identification of mislabels. By involving the neural network's confidence and predictions of a GMM, presumably noisy labels were

corrected or filtered as validated on the two largest VCE datasets containing pathologies. 100 data samples were re-evaluated by experienced gastroenterologists, which validated the effectiveness of the presented approach. This resulted in cleaned datasets usable for final model development. It was demonstrated that prior dataset cleaning indeed enhances image classification results by ML-based models. Compared to current baselines, the anomaly detection results were significantly boosted. One drawback of the applied method was that classes with smaller sample sizes were more frequently filtered out or falsely corrected than other classes. Since the neural network sees the smaller classes less often during the training process, they produce a higher loss. This loss then functions as a basis for the correction. To protect the underrepresented data, one GMM could be used for each class, such that outliers are only detected within the classes. Varying the number of GMM components could also enhance the framework's performance, since a larger number of components might be more flexible in detecting label noise. To conclude, it is advisable to review all identified data samples by clinicians and provide updated annotations. Nevertheless, the provided cleaned dataset as produced by the introduced framework itself provides a substantial improvement for the anomaly detection task for the VCE.

Apart from that, underrepresentation of labeled data was addressed by using autoencoders to leverage the vast amount of unlabeled data within this thesis. Different methods were successfully employed to lower the impact of the present class imbalances. To exclude potential domain shifts regarding data acquisition and measurement, the seizure detection pipeline can be verified using data obtained from other hospitals in the future. For the VCE experiments, a domain shift was counteracted by incorporating the Galar dataset, which did not only include patient studies captured with different image resolutions, but also contains studies from different hospitals and data acquired with different VCE capsules from multiple manufacturers. In the future, it is recommended to combine the ensemble-based anomaly detection approach with the cleaned dataset. It can be expected that this further boosts the classification performance.

While quantization positively impacts the computational and memory costs, it also imposes a challenge due to the lowered accompanied precision, which was constantly considered within this process. Hardware efficiency and superior classification performance was demonstrated for the localization task of the VCE saving 90% of energy on average before the small intestine was entered in comparison to standard VCE procedures outperforming current baselines. The EEG-based seizure detection approach involving the UltraTrail hardware accelerator features a substantial enhancement to current neuroimplants by outperforming current baselines, obtaining a power consumption of only 495 nW and an energy demand of 49.5 µJ. In this experimental setup, the largest fraction of the overall power consumption was required by the control unit. Notably, this was reduced in a following work by improving the generation of memory addresses [Pal+25]. In the future, the seizure detection approach can be optimized further to rely on a smaller amount of EEG electrodes for data acquisition, leading to a reduced power consumption and slightly lower on-chip area due to a smaller model. For final evaluations, the accelerator should be included in a full prototype to assess the energy demand of a whole seizure detection system. The anomaly detection requires further refinement to boost its sensitivity before conducting more experiments on the necessary energy demand. As has been demonstrated regarding the problems of VCE localization and the seizure detection in this work, for the concluding hardware simulation of the anomaly detection, the model size must be considered. This influences the limited on-chip area, along with using only practical operations supported by the hardware and the strict avoidance of floating-point representation in the final model.

An alternative to computing the Viterbi decoding directly on the core is to program an accelerator, which is specifically tailored for this application. It can be assumed that such accelerator requires fewer clock cycles than the computation on the core. Besides the number of clock cycles, one has to consider the required on-chip area of an additional accelerator and whether a core might be needed for other tasks anyway. If the accelerator performs better and the core is only needed for Viterbi decoding, then it is reasonable to use such accelerator instead. However, the core can be used for less specific operations as well as other tasks in contrast to the accelerator. Furthermore, additional area needs to be reserved for an accelerator. Hence, using only a core for computing this post-processing might be more practical. Such trade-offs must be ultimately chosen by hardware design specialists.

Within this thesis, lightweight classification models were designed and validated using ultra low-power hardware architectures which outperform current baselines, that rely on complex structures, such as vision transformers, for the given medical applications. The proposed techniques are hardware-efficient and thoughtful integrated into existing systems, require only minimal power and area, as well as further reduce the overall energy demand.

Chapter 4

Conclusion and Outlook

4.1 Conclusion

Given the necessity of embedding useful AI models into in-body sensor edge devices in medical applications, this thesis explored hardware-aware machine learning methods for targeting such low-power edge devices to improve the overall procedures and enable AI-based decision-making in the future. This was demonstrated with two medical applications: the Video Capsule Endoscopy (VCE) for the investigation of the small intestine and seizure detection using neuroimplants for drug-resistant epilepsy patients. In both cases, the main research objective was to prolong the battery lifetime of such embedded sensor edge devices while providing a competitive classification performance. Machine learning methods were conceptualized while addressing the challenges imposed by the resource-restricted devices resulting in stringent requirements on models targeting applications involving low-power sensor edge devices. As presented and discussed in Chapter 3 the experiments and publications fulfill the research objectives, which are once again listed in the following.

Research Objectives

- ✓ Accurate hardware-aware EEG-based seizure detection pipeline (Publication A1, Section 3.1.1)
- ✓ Demonstrating an ultra low energy consumption of the seizure detection approach by hardware simulation (Publication A1, Section 3.1.1).
- ✓ Designing a hardware-aware ML model performing precise organ detection in VCE (Publication A2, Section 3.2.1; Publication A6, Section 3.2.5).
- ✓ Substantially enhancing anomaly detection in VCE studies (Publication A3, Section 3.2.2; Publication A4, Section 3.2.3; Publication A5, Section 3.2.4).
- ✓ Realizing joint localization and anomaly detection (Publication A4, Section 3.2.3).
- ✓ Saving energy by omitting unnecessary image preprocessing steps and model simulation on a simple VCE demonstrator (Publication A6, Section 3.2.5).
- ✓ Prolonging the battery lifetime of a VCE device based on AI-based decision models (Publication A6; Section 3.2.5).

For the seizure detection, a hardware-aware EEG-based seizure detection approach was presented and simulated with the UltraTrail hardware accelerator, leading to a significant reduction in overall power and energy consumption compared to literature baselines, which

potentially prolongs the battery lifetime of such implants in the future by exhibiting an energy demand of only 49.5 µJ. Inclusion of time-series analysis for subsequent post-processing further improved the classification results notably. Furthermore, instead of 16 channels, experiments using only 8, 4 and 2 channels demonstrated that a channel reduction is practical with the presented approach, and can be used for further optimization, fine-tuning and hardware simulation. A lower number of utilized channels reduces the input feature map and is expected to result in a smaller on-chip area.

In the VCE field, hardware-aware ML models for precise organ detection and localization within the GI tract competing with current baselines were presented, as well as an ensemble learning strategy to significantly enhance the anomaly detection in the GI tract based on VCE images. The anomaly detection approach was then verified on the two largest publicly available VCE datasets. Combining both tasks, a multi-task model capable of anatomical localization and anomaly detection in a single model was designed. Additionally, a framework for the identification, correction and filtering of presumably noisy labels was designed and used to generate cleaned versions of the available VCE datasets comprising pathologies. Training a neural network on such cleaned dataset yielded a substantial enhancement regarding the classification performance. Finally, the training of neural networks directly on images in raw Bayer pattern format was tested to save the formerly necessary conversion step to RGB images before classification. The accurate determination of the capsule's stomach exit permits image transmission only after entering the small intestine which saves a notable amount of energy by eliminating the necessity of sending out images beforehand. This was simulated on a simple VCE hardware demonstrator which yielded a prolonged battery lifetime due to the usage of an AI-based decision model in comparison to capsules without ML models integrated.

The results represent important key contributions for personalized AI-based medicine targeting edge devices. While this work focused on the two applications of neuroimplants for seizure detection and the VCE, the described methods can also be applied in other contexts and can function as a starting point for future work.

4.2 Outlook

This thesis lays the foundation for hardware-aware machine learning methods providing new functionalities targeting medical edge devices and offers various ideas for further improvement. For both applications, these potential refinements are outlined in the following. If neuroimplants are equipped with an EEG-based classification model, the number of electrodes and EEG channels will be minimized as much as possible. Thus, focusing on a limited number of EEG channels is essential. The Viterbi decoding can be enhanced as proposed in Section 3.1, by incorporating a larger number of emissions in the HMM. This mitigates the accompanied information loss of transferring the data information from the CNN to the HMM and potentially improves the overall classification performance further. While usually only two classes (ictal, non-ictal) are available in most EEG datasets, introducing a third class named preictal, by labeling a defined period before a seizure start, provides valuable information and enables to extend the HMM approach by a third class. This combinatorial approach also relies on the prediction performance of the CNN, which requires a substantial amount of data for training. Since the preictal data is limited by the total number of seizures within a dataset, multiple datasets need to be merged or a larger dataset than the CHB-MIT database must be used to enable effective neural network training. If an improved Viterbi approach yields a superior performance, this can also facilitate an additional channel reduction, leading to a decreased energy demand.

Another procedure which might benefit from the presented techniques is capturing rhythmic abnormalities within the heart based on data from electrocardiographies, capturing the electrical activity of the heart. Similarly to the EEG-based seizure detection, a CNN can be combined with a HMM and Viterbi decoding, leveraging the time-series data to screen for anomalies. Considering that such devices must also operate in real-time and require low-power hardware architectures, the presented methods are well-suited for such tasks.

Considering the effective combination of Viterbi decoding and CNNs for the VCE, further refinement in this matter might enhance the classification performance even further. For example, the Hidden Markov Model and subsequent Viterbi decoding can be extended with time-dependent transitions. More specifically, this would only permit transitioning to the next organ after a determined time has passed, which potentially reduces the number of too early detections of entering the small intestine. Additionally, computing the log-likelihood matrix fully during Viterbi decoding after a transition has been detected might be beneficial.

In the current implementation, the starting probabilities are chosen so that the algorithm generally starts in the first organ. Instead, in the case of a detected transition, the starting probabilities might be updated so that the current organ has the highest probability of being the correct state. Furthermore, as an alternative to parsing every image from the test set to the Viterbi decoding, one could extend this implementation, so that only every n^{th} image is evaluated to simulate a reduced frame rate. Enhancing the organ classification further, a cross-evaluation on all available VCE datasets containing annotations with organ labels should be conducted. This could be realized by training a single CNN with all available datasets and test the same model on all test sets individually to generate a more robust CNN. By conducting accurate localization of the capsule in real-time, additional features can be employed on-device in the future. For example, releasing medication at a specific location, adapting the frame rate depending on the interest at a specific region or zooming in and out features.

Building up on the low-power consumption obtained with the low-complex CNN used for the localization task, the hardware can be adjusted to simulate the energy demand of the multi-task model. Additionally, the multi-task model might be extended by a third or even more tasks. For example, besides providing organ classification and anomaly detection capabilities, it could be further expanded to conduct visibility assessment. Further optimization of the neural network by hardware-software co-design using a neural architecture search is advisable to minimize the model size optimized for the target hardware while having a remarkable classification performance. To additionally enhance the model performance, this can be integrated with subsequent Viterbi decoding. Considering that the largest area on the hardware accelerator is still occupied by the storage of the weights of a neural network, in the future, weight and feature compression of the employed CNN is recommendable. Furthermore, implementing image compression on VCE devices can lower the area demand.

In addition, the cleaning pipeline could be further improved by replacing the GMM with other mixture models to investigate if they can depict underlying loss distributions better. Protecting classes with fewer samples by adapting the GMM approach, for example by using more GMMs or a varying number of GMM components, is further advisable. In the future, the ensemble model approach to perform anomaly detection should be applied to the cleaned VCE datasets. If possible, the identified noisy samples should be re-annotated by gastroenterologists. Combining the presented ensemble approach with the cleaned

datasets is expected to further enhance the detection of pathologies on low-resolution VCE images.

To conclude, by focusing on the application aspect and additionally bearing the stringent hardware constraints in mind, this thesis represents a crucial contribution to the usage of AI in medical sensor in-body edge devices, offering personalized and effective screening as well as disease detection.

References

[Abd+14] Ossama Abdel-Hamid et al. “Convolutional neural networks for speech recognition.” In: *IEEE/ACM Transactions on audio, speech, and language processing* 22.10 (2014), pp. 1533–1545.

[Abi+25] Arefin Ittesafun Abian et al. “Atrous spatial pyramid pooling with swin transformer model for classification of gastrointestinal tract diseases from videos with enhanced explainability.” In: *Engineering Applications of Artificial Intelligence* 150 (2025), p. 110656.

[AC15] Jinwon An and Sungzoon Cho. “Variational autoencoder based anomaly detection using reconstruction probability.” In: *Special lecture on IE* 2.1 (2015), pp. 1–18.

[AG24] ams-OSRAM AG. *NanEyeC Miniature Camera Module*. DS000503. v5-00, https://look.ams-osram.com/m/19863b4335e67f1b/original/NanEyeC-Miniature-Camera-Module.pdf. Oct. 2024.

[Ail+15] Pierre Ailliot et al. “Stochastic weather generators: an overview of weather type models.” In: *Journal de la société française de statistique* 156.1 (2015), pp. 101–113.

[ALN20] Séverine Affeldt, Lazhar Labiod, and Mohamed Nadif. “Spectral clustering via ensemble deep autoencoder learning (SC-EDAE).” In: *Pattern Recognition* 108 (2020), p. 107522.

[And+01] Ralph G Andrzejak et al. “Indications of nonlinear deterministic and finite-dimensional structures in time series of brain electrical activity: Dependence on recording region and brain state.” In: *Physical Review E* 64.6 (2001), p. 061907.

[Aok+19] Tomonori Aoki et al. “Automatic detection of erosions and ulcerations in wireless capsule endoscopy images based on a deep convolutional neural network.” In: *Gastrointestinal endoscopy* 89.2 (2019), pp. 357–363.

[Arn+21] Anurag Arnab et al. “Vivit: A video vision transformer.” In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2021, pp. 6836–6846.

[Aur+04] H Aurlien et al. “EEG background activity described by a large computerized database.” In: *Clinical Neurophysiology* 115.3 (2004), pp. 665–673.

[Bah+21] Andreas Bahr et al. “Epileptic seizure detection on an ultra-low-power embedded RISC-V processor using a convolutional neural network.” In: *Biosensors* 11.7 (2021), p. 203.

[Bao+12] Guanqun Bao et al. “Modeling of the movement of the endoscopy capsule inside gi tract based on the captured endoscopic images.” In: *Proceedings of the IEEE International Conference on Modeling, Simulation and Visualization Methods, MSV*. Vol. 12. 2012.

[Bao+14] Guanqun Bao et al. “A computer vision based speed estimation technique for localiz ing the wireless capsule endoscope inside small intestine.” In: *36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. Vol. 123. 2014.

[Bau+25] Oliver Bause et al. “Smart Video Capsule Endoscopy: Raw Image-Based Localization for Enhanced GI Tract Investigation.” In: *International Conference on Neural Information Processing*. Springer. 2025, pp. 33–47.

[BBA20] Amir Jalaly Bidgoly, Hamed Jalaly Bidgoly, and Zeynab Arezoumand. “A survey on methods and challenges in EEG based authentication.” In: *Computers & Security* 93 (2020), p. 101788.

[Ben+21] Hadjer Benmeziane et al. “A comprehensive survey on hardware-aware neural architecture search.” In: *arXiv preprint arXiv:2101.09336* (2021).

[Ben02] Elinor Ben-Menachem. “Vagus-nerve stimulation for the treatment of epilepsy.” In: *The Lancet Neurology* 1.8 (2002), pp. 477–482.

[Ber+20] Paul Palomero Bernardo et al. “Ultratrail: A configurable ultralow-power tc-resnet ai accelerator for efficient keyword spotting.” In: *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems* 39.11 (2020), pp. 4240–4251.

[Ber29] Hans Berger. “Über das elektroenkephalogramm des menschen.” In: *Archiv für psychiatrie und nervenkrankheiten* 87.1 (1929), pp. 527–570.

[BGV92] Bernhard E Boser, Isabelle M Guyon, and Vladimir N Vapnik. “A training algorithm for optimal margin classifiers.” In: *Proceedings of the fifth annual workshop on Computational learning theory*. 1992, pp. 144–152.

[Bjo+18] Nils Bjorck et al. “Understanding batch normalization.” In: *Advances in neural information processing systems* 31 (2018).

[BK88] Hervé Bourlard and Yves Kamp. “Auto-association by multilayer perceptrons and singular value decomposition.” In: *Biological cybernetics* 59.4 (1988), pp. 291–294.

[BLB22] Oliver Bringmann, Walter Lange, and Martin Bogdan. *Eingebettete Systeme: Entwurf, Synthese und Edge AI*. Walter de Gruyter GmbH & Co KG, 2022.

[BMP13] Guanqun Bao, Liang Mi, and Kaveh Pahlavan. “A video aided RF localization technique for the wireless capsule endoscope (WCE) inside small intestine.” In: *Proceedings of the 8th International Conference on Body Area Networks*. 2013, pp. 55–61.

[BP66] Leonard E Baum and Ted Petrie. “Statistical inference for probabilistic functions of finite state Markov chains.” In: *The annals of mathematical statistics* 37.6 (1966), pp. 1554–1563.

[Bre01] Leo Breiman. “Random forests.” In: *Machine learning* 45 (2001), pp. 5–32.

[Bre02] Leo Breiman. “Manual on setting up, using, and understanding random forests v3. 1.” In: *Statistics Department University of California Berkeley, CA, USA* 1.58 (2002), pp. 3–42.

[Bre96] Leo Breiman. “Bagging predictors.” In: *Machine learning* 24 (1996), pp. 123–140.

[Car93] R Caruana. “Multitask learning: A knowledge-based source of inductive bias1.” In: *Proceedings of the Tenth International Conference on Machine Learning*. 1993, pp. 41–48.

[CBK09] Varun Chandola, Arindam Banerjee, and Vipin Kumar. “Anomaly detection: A survey.” In: *ACM computing surveys (CSUR)* 41.3 (2009), pp. 1–58.

[CG16] Tianqi Chen and Carlos Guestrin. “Xgboost: A scalable tree boosting system.” In: *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. 2016, pp. 785–794.

[Cha+22] Amber Charoen et al. “Rhode Island gastroenterology video capsule endoscopy data set.” In: *Scientific Data* 9.1 (2022), p. 602.

[Cho+19] Seungwoo Choi et al. “Temporal convolution for real-time keyword spotting on mobile devices.” In: *arXiv preprint arXiv:1904.03814* (2019).

[Cho17] François Chollet. "Xception: Deep learning with depthwise separable convolutions." In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 1251–1258.

[CHP19] David R Cave, Shahrad Hakimian, and Krunal Patel. "Current controversies concerning capsule endoscopy." In: *Digestive diseases and sciences* 64.11 (2019), pp. 3040–3047.

[Chu+23] Joowon Chung et al. "Automatic classification of GI organs in wireless capsule endoscopy using a no-code platform-based deep learning model." In: *Diagnostics* 13.8 (2023), p. 1389.

[Chu89] Gary A Churchill. "Stochastic models for heterogeneous DNA sequences." In: *Bulletin of mathematical biology* 51.1 (1989), pp. 79–94.

[CL21] Mahesh Chandra and Brejesh Lall. "A novel method for cnn training using existing color datasets for classifying hand postures in bayer images." In: *SN Computer Science* 2.2 (2021), p. 60.

[Cob+58] W Cobb et al. "Report of the committee on methods of clinical examination in electroencephalography." In: *Electroencephalogr. Clin. Neurophysiol* 10.2 (1958), pp. 370–375.

[CPC16] Alfredo Canziani, Adam Paszke, and Eugenio Culurciello. "An analysis of deep neural network models for practical applications." In: *arXiv preprint arXiv:1605.07678* (2016).

[CV95] Corinna Cortes and Vladimir Vapnik. "Support-vector networks." In: *Machine learning* 20 (1995), pp. 273–297.

[De +11] Jane De Tisi et al. "The long-term outcome of adult epilepsy surgery, patterns of seizure remission, and relapse: a cohort study." In: *The Lancet* 378.9800 (2011), pp. 1388–1395.

[Die00] Thomas G Dietterich. "Ensemble methods in machine learning." In: *International workshop on multiple classifier systems*. Springer. 2000, pp. 1–15.

[DK20] Deba Prasad Dash and Maheshkumar H Kolekar. "Hidden Markov model based epileptic seizure detection using tunable Q wavelet transform." In: *Journal of biomedical research* 34.3 (2020), p. 170.

[Doc] PyTorch Documentation2025. *Introduction to PyTorch Tensors*. URL: https://docs.pytorch.org/tutorials/beginner/introyt/tensors_deeper_tutorial.html (visited on 08/22/2025).

[DP15] Edward S Dove and Mark Phillips. "Privacy law, data sharing policies, and medical data: a comparative perspective." In: *Medical data privacy handbook* (2015), pp. 639–678.

[Dun+06] John S Duncan et al. "Adult epilepsy." In: *The Lancet* 367.9516 (2006), pp. 1087–1100.

[Dur+98] Richard Durbin et al. *Biological sequence analysis: probabilistic models of proteins and nucleic acids*. Cambridge university press, 1998.

[Ech+99] J Echauz et al. "Median-based filtering methods for EEG seizure detection." In: *Proceedings of the First Joint BMES/EMBS Conference. 1999 IEEE Engineering in Medicine and Biology 21st Annual Conference and the 1999 Annual Fall Meeting of the Biomedical Engineering Society (Cat. N)*. Vol. 1. IEEE. 1999, 439–vol.

[Edd96] Sean R Eddy. "Hidden markov models." In: *Current opinion in structural biology* 6.3 (1996), pp. 361–365.

[Ein+23] Athar A Ein Shoka et al. "EEG seizure detection: concepts, techniques, challenges, and future trends." In: *Multimedia Tools and Applications* 82.27 (2023), pp. 42021–42051.

[Est+01] Rosana Esteller et al. "Line length: an efficient feature for seizure onset detection." In: *2001 Conference Proceedings of the 23rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. Vol. 2. IEEE. 2001, pp. 1707–1710.

[Est+19] Andre Esteva et al. "A guide to deep learning in healthcare." In: *Nature medicine* 25.1 (2019), pp. 24–29.

[Far+21] Nuruzzaman Faruqui et al. "LungNet: A hybrid deep-CNN model for lung cancer diagnosis using CT and wearable sensor-based medical IoT data." In: *Computers in Biology and Medicine* 139 (2021), p. 104961.

[FG22] Yuming Fu and Yong-Xin Guo. "Wearable permanent magnet tracking system for wireless capsule endoscope." In: *IEEE Sensors Journal* 22.8 (2022), pp. 8113–8122.

[Fir+03] Z Fireman et al. "Diagnosing small bowel Crohn's disease with wireless capsule endoscopy." In: *Gut* 52.3 (2003), pp. 390–392.

[Fis+04] Doron Fischer et al. "Capsule endoscopy: the localization system." In: *Gastrointestinal Endoscopy Clinics* 14.1 (2004), pp. 25–31.

[Fis+14] Robert S Fisher et al. "ILAE official report: a practical clinical definition of epilepsy." In: *Epilepsia* 55.4 (2014), pp. 475–482.

[Fly72] Michael J Flynn. "Some computer organizations and their effectiveness." In: *IEEE transactions on computers* 100.9 (1972), pp. 948–960.

[For73] G David Forney. "The viterbi algorithm." In: *Proceedings of the IEEE* 61.3 (1973), pp. 268–278.

[FQD20] Wei Feng, Yinghui Quan, and Gabriel Dauphin. "Label noise cleaning with an adaptive ensemble method based on noise detection metric." In: *Sensors* 20.23 (2020), p. 6718.

[Fri+18] Maayan Frid-Adar et al. "GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification." In: *Neurocomputing* 321 (2018), pp. 321–331.

[Gal98] Mark JF Gales. "Maximum likelihood linear transformations for HMM-based speech recognition." In: *Computer speech & language* 12.2 (1998), pp. 75–98.

[GBC16] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. http://www.deeplearningbook.org. MIT Press, 2016.

[Gon+19] Dong Gong et al. "Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection." In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2019, pp. 1705–1714.

[Got82] Jean Gotman. "Automatic recognition of epileptic seizures in the EEG." In: *Electroencephalography and clinical Neurophysiology* 54.5 (1982), pp. 530–540.

[GS12] Charles Miller Grinstead and James Laurie Snell. *Introduction to probability*. American Mathematical Soc., 2012.

[Guo+19] Yunhui Guo et al. "Depthwise convolution is all you need for learning multiple visual domains." In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 33. 01. 2019, pp. 8368–8375.

[Gup+15] Suyog Gupta et al. "Deep learning with limited numerical precision." In: *International conference on machine learning*. PMLR. 2015, pp. 1737–1746.

[Gut10] John Guttag. "CHB-MIT Scalp EEG Database." In: *PhysioNet* (June 2010). Version 1.0.0. DOI: 10.13026/C2K01R. URL: https://doi.org/10.13026/C2K01R.

[Hec+14] Christianne N Heck et al. "Two-year seizure reduction in adults with medically intractable partial onset epilepsy treated with responsive neurostimulation: final results of the RNS System Pivotal trial." In: *Epilepsia* 55.3 (2014), pp. 432–441.

[Ho95] Tin Kam Ho. "Random decision forests." In: *Proceedings of 3rd international conference on document analysis and recognition*. Vol. 1. IEEE. 1995, pp. 278–282.

[Hol+20] Olle G Holmberg et al. "Self-supervised retinal thickness prediction enables deep learning from unlabelled data to boost classification of diabetic retinopathy." In: *Nature Machine Intelligence* 2.11 (2020), pp. 719–726.

[How+17] Andrew G Howard et al. "Mobilenets: Efficient convolutional neural networks for mobile vision applications." In: *arXiv preprint arXiv:1704.04861* (2017).

[How+19] Andrew Howard et al. "Searching for mobilenetv3." In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2019, pp. 1314–1324.

[HP11] John L Hennessy and David A Patterson. *Computer architecture: a quantitative approach*. Elsevier, 2011.

[HSS18] Jie Hu, Li Shen, and Gang Sun. "Squeeze-and-excitation networks." In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 7132–7141.

[Hüg+18] Maria Hügle et al. "Early seizure detection with an energy-efficient convolutional neural network on an implantable microcontroller." In: *2018 International Joint Conference on Neural Networks (IJCNN)*. IEEE. 2018, pp. 1–7.

[Iak+13] Dimitris K Iakovidis et al. "Capsule endoscope localization based on visual features." In: *13th IEEE International Conference on BioInformatics and BioEngineering*. IEEE. 2013, pp. 1–4.

[Idd+00] Gavriel Iddan et al. "Wireless capsule endoscopy." In: *Nature* 405.6785 (2000), pp. 417–417.

[Ihl+12] Matthias Ihle et al. "EPILEPSIAE–A European epilepsy database." In: *Computer methods and programs in biomedicine* 106.3 (2012), pp. 127–138.

[IS04] Gavriel J Iddan and C Paul Swain. "History and development of capsule endoscopy." In: *Gastrointestinal Endoscopy Clinics* 14.1 (2004), pp. 1–9.

[IS15] Sergey Ioffe and Christian Szegedy. "Batch normalization: Accelerating deep network training by reducing internal covariate shift." In: *International conference on machine learning*. pmlr. 2015, pp. 448–456.

[Isl+24] Saidul Islam et al. "A comprehensive survey on applications of transformers for deep learning tasks." In: *Expert Systems with Applications* 241 (2024), p. 122666.

[IZM] Copyright Volker Mai / Fraunhofer IZM. *Schnellere Dünndarm-Diagnose dank Kamerapille*. URL: https://www.izm.fraunhofer.de/de/news_events/tech_news/duenndarm_diagnose_dank_kamerapille.html (visited on 02/28/2025).

[Jai+21] Samir Jain et al. "A deep CNN model for anomaly detection and localization in wireless capsule endoscopy images." In: *Computers in Biology and Medicine* 137 (2021), p. 104789.

[Jha+20] Debesh Jha et al. "Doubleu-net: A deep convolutional neural network for medical image segmentation." In: *2020 IEEE 33rd International symposium on computer-based medical systems (CBMS)*. IEEE. 2020, pp. 558–564.

[Jia+24] Gaoxia Jiang et al. "Which is more effective in label noise cleaning, correction or filtering?" In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 38. 2024, pp. 12866–12873.

[JMM21] Imene Jemal, Amar Mitiche, and Neila Mezghani. "A Study of EEG Feature Complexity in Epileptic Seizure Prediction." In: *Applied Sciences* 11.4 (2021). ISSN: 2076-3417. DOI: 10.3390/app11041579. URL: https://www.mdpi.com/2076-3417/11/4/1579.

[JSA20] Gopal Chandra Jana, Ratna Sharma, and Anupam Agrawal. “A 1D-CNN-spectrogram based approach for seizure detection from EEG signal.” In: *Procedia Computer Science* 167 (2020), pp. 403–412.

[JV23] Vikram Jain and Marian Verhelst. *Towards Heterogeneous Multi-core Systems-on-Chip for Edge Machine Learning: Journey from Single-core Acceleration to Multi-core Heterogeneous Systems*. Springer Nature, 2023.

[Kan+18] Xudong Kang et al. “Detection and correction of mislabeled training samples for hyperspectral image classification.” In: *IEEE Transactions on Geoscience and Remote Sensing* 56.10 (2018), pp. 5673–5686.

[Kar+20] Davood Karimi et al. “Deep learning with noisy labels: Exploring techniques and remedies in medical image analysis.” In: *Medical image analysis* 65 (2020), p. 101759.

[KGC18] Alex Kendall, Yarin Gal, and Roberto Cipolla. “Multi-task learning using uncertainty to weigh losses for scene geometry and semantics.” In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 7482–7491.

[KH+09] Alex Krizhevsky, Geoffrey Hinton, et al. “Learning multiple layers of features from tiny images.” In: (2009).

[Kha+21] Alaa O Khadidos et al. “Bayer image demosaicking and denoising based on specialized networks using deep learning.” In: *Multimedia Systems* 27.4 (2021), pp. 807–819.

[Kir+18] Isabell Kiral-Kornek et al. “Epileptic seizure prediction using big data and deep learning: toward a mobile system.” In: *EBioMedicine* 27 (2018), pp. 103–111.

[Kon+21] Zishang Kong et al. “Multi-task classification and segmentation for explicable capsule endoscopy diagnostics.” In: *Frontiers in Molecular Biosciences* 8 (2021), p. 614277.

[KPK23] SP Predeep Kumar, K John Peter, and C Sahaya Kingsly. “De-noising and Demosaicking of Bayer image using deep convolutional attention residual learning.” In: *Multimedia Tools and Applications* 82.13 (2023), pp. 20323–20342.

[KS20] Kristina Kravalis and Andreas Schulze-Bonhage. “PIMIDES I: a pilot study to assess the feasibility of patient-controlled neurostimulation with the EASEE® system to treat medically refractory focal epilepsy.” In: *Neurological Research and Practice* 2 (2020), pp. 1–3.

[KSH12] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. “Imagenet classification with deep convolutional neural networks.” In: *Advances in neural information processing systems* 25 (2012).

[KW03] Ludmila I Kuncheva and Christopher J Whitaker. “Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy.” In: *Machine learning* 51 (2003), pp. 181–207.

[Lan02] Douglas J Lanska. “JL Corning and vagal nerve stimulation for seizures in the 1880s.” In: *Neurology* 58.3 (2002), pp. 452–459.

[Lau14] Sifre Laurent. “Rigid-motion scattering for image classification.” In: *Ph. D. thesis section* 6.2 (2014).

[LB+95] Yann LeCun, Yoshua Bengio, et al. “Convolutional networks for images, speech, and time series.” In: *The handbook of brain theory and neural networks* 3361.10 (1995), p. 1995.

[Le +25] Maxime Le Floch et al. “Galar-a large multi-label video capsule endoscopy dataset.” In: *Scientific Data* 12.1 (2025), p. 828.

[LeC98] Yann LeCun. "The MNIST database of handwritten digits." In: *http://yann. lecun. com/exdb/mnist/* (1998).

[Lee+07] Jeongkyu Lee et al. "Automatic classification of digestive organs in wireless capsule endoscopy videos." In: *Proceedings of the 2007 ACM symposium on Applied computing*. 2007, pp. 1041–1045.

[Lee+13] Dong-Hyun Lee et al. "Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks." In: *Workshop on challenges in representation learning, ICML*. Vol. 3. 2. Atlanta. 2013, p. 896.

[Lee+18] Kuang-Huei Lee et al. "Cleannet: Transfer learning for scalable image classifier training with label noise." In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 5447–5456.

[LF87] Yann Le Cun and Françoise Fogelman-Soulié. "Modèles connexionnistes de l'apprentissage." In: *Intellectica* 2.1 (1987), pp. 114–143.

[Lin+17] Tsung-Yi Lin et al. "Focal loss for dense object detection." In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 2980–2988.

[Lit+01] Brian Litt et al. "Epileptic seizures may begin hours in advance of clinical onset: a report of five patients." In: *Neuron* 30.1 (2001), pp. 51–64.

[Liu+21] Ze Liu et al. "Swin transformer: Hierarchical vision transformer using shifted windows." In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2021, pp. 10012–10022.

[LJD19] Shikun Liu, Edward Johns, and Andrew J Davison. "End-to-end multi-task learning with attention." In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019, pp. 1871–1880.

[Loz+19] Andres M Lozano et al. "Deep brain stimulation: current challenges and future directions." In: *Nature Reviews Neurology* 15.3 (2019), pp. 148–160.

[LSH20] Junnan Li, Richard Socher, and Steven CH Hoi. "Dividemix: Learning with noisy labels as semi-supervised learning." In: *arXiv preprint arXiv:2002.07394* (2020).

[Lut+19] Brendon Lutnick et al. "An integrated iterative annotation technique for easing neural network training in medical image analysis." In: *Nature machine intelligence* 1.2 (2019), pp. 112–119.

[Man+22] Farrokh Manzouri et al. "A comparison of energy-efficient seizure detectors for implantable neurostimulation devices." In: *Frontiers in Neurology* 12 (2022), p. 703797.

[Mar+14] Neil Marya et al. "Computerized 3-dimensional localization of a video capsule in the abdominal cavity: validation by digital radiography." In: *Gastrointestinal endoscopy* 79.4 (2014), pp. 669–674.

[Mat+17] Bertrand Mathon et al. "Predictive factors of long-term outcomes of surgery for mesial temporal lobe epilepsy associated with hippocampal sclerosis." In: *Epilepsia* 58.8 (2017), pp. 1473–1485.

[MC14] Negin Moghim and David W Corne. "Predicting epileptic seizures in advance." In: *PloS one* 9.6 (2014), e99334.

[Med] Medtronic. *PillCam™ SB 3 capsule*. URL: https://www.medtronic.com/en-us/healthcare-professionals/products/digestive-gastrointestinal/capsule-endoscopy/capsules/pillcam-sb-3-capsule.html (visited on 02/28/2025).

[Min+21] Shervin Minaee et al. "Image segmentation using deep learning: A survey." In: *IEEE transactions on pattern analysis and machine intelligence* 44.7 (2021), pp. 3523–3542.

[Mon+16] Sara Monteiro et al. "PillCam® SB3 capsule: Does the increased frame rate eliminate the risk of missing lesions?" In: *World journal of gastroenterology* 22.10 (2016), p. 3066.

[Mor+13] George L Morris III et al. “Evidence-based guideline update: vagus nerve stimulation for the treatment of epilepsy: report of the Guideline Development Subcommittee of the American Academy of Neurology.” In: *Neurology* 81.16 (2013), pp. 1453–1459.

[MP00] Geoffrey J McLachlan and David Peel. *Finite mixture models*. John Wiley & Sons, 2000.

[Nas+18] Sheraz Naseer et al. “Enhanced network anomaly detection based on deep neural networks.” In: *IEEE access* 6 (2018), pp. 48231–48246.

[OM99] David Opitz and Richard Maclin. “Popular ensemble methods: An empirical study.” In: *Journal of artificial intelligence research* 11 (1999), pp. 169–198.

[OP16] Iyad Obeid and Joseph Picone. “The temple university hospital EEG data corpus.” In: *Frontiers in neuroscience* 10 (2016), p. 196.

[Ost+18] Pavel Ostyakov et al. “Label denoising with large ensembles of heterogeneous neural networks.” In: *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*. 2018, pp. 0–0.

[PA14] Duc Minh Pham and Syed Mahfuzul Aziz. “A real-time localization system for an endoscopic capsule using magnetic sensors.” In: *Sensors* 14.11 (2014), pp. 20910–20929.

[Pah+12] Kaveh Pahlavan et al. “RF localization for wireless video capsule endoscopy.” In: *International Journal of Wireless Information Networks* 19 (2012), pp. 326–340.

[Pal+25] Paul Palomero Bernardo et al. “Compiler-aware AI Hardware Design for Edge Devices.” In: *Proceedings of the 8th International Workshop on Edge Systems, Analytics and Networking*. 2025, pp. 31–36.

[Pat+17] Giorgio Patrini et al. “Making deep neural networks robust to label noise: A loss correction approach.” In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 1944–1952.

[PDG13] Piero Perucca, François Dubeau, and Jean Gotman. “Widespread EEG changes precede focal seizures.” In: *PloS one* 8.11 (2013), e80972.

[Pen+04] Marco Pennazio et al. “Outcome of patients with obscure gastrointestinal bleeding after capsule endoscopy: report of 100 consecutive cases.” In: *Gastroenterology* 126.3 (2004), pp. 643–653.

[Pet+01] Thomas E Peters et al. “Network system for automated seizure detection and contingent delivery of therapy.” In: *Journal of clinical neurophysiology* 18.6 (2001), pp. 545–549.

[PFJ06] Haim Permuter, Joseph Francos, and Ian Jermyn. “A study of Gaussian mixture models of color and texture features for image classification and segmentation.” In: *Pattern recognition* 39.4 (2006), pp. 695–706.

[PH16] David A Patterson and John L Hennessy. *Computer organization and design ARM edition: the hardware software interface*. Morgan kaufmann, 2016.

[Pre] Precisis. *EASEE®*. URL: https://easee.precisis.de/de/ (visited on 04/06/2025).

[Rab89] Lawrence R Rabiner. “A tutorial on hidden Markov models and selected applications in speech recognition.” In: *Proceedings of the IEEE* 77.2 (1989), pp. 257–286.

[Ras+20] Mamunur Rashid et al. “Current status, challenges, and possible solutions of EEG-based brain-computer interface: a comprehensive review.” In: *Frontiers in neurorobotics* 14 (2020), p. 25.

[RCR14] Philippe Ryvlin, J Helen Cross, and Sylvain Rheims. “Epilepsy surgery in children and adults.” In: *The Lancet Neurology* 13.11 (2014), pp. 1114–1126.

[Reg+25] Smriti Regmi et al. "Vision transformer for efficient chest X-ray and gastrointestinal image classification." In: *Medical Imaging 2025: Computer-Aided Diagnosis*. Vol. 13407. SPIE. 2025, pp. 912–923.

[Rem+12] Michael Remmert et al. "HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment." In: *Nature methods* 9.2 (2012), pp. 173–175.

[Ren+94] Steve Renals et al. "Connectionist probability estimators in HMM speech recognition." In: *IEEE transactions on speech and audio processing* 2.1 (1994), pp. 161–174.

[RHW85] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. *Learning internal representations by error propagation*. Tech. rep. 1985.

[RI17] Glenna L Read and Isaiah J Innis. "Electroencephalography (eeg)." In: *The international encyclopedia of communication research methods* (2017), pp. 1–18.

[Riz+21] Mamshad Nayeem Rizve et al. "In defense of pseudo-labeling: An uncertainty-aware pseudo-label selection framework for semi-supervised learning." In: *arXiv preprint arXiv:2101.06329* (2021).

[RMK16] Georgia Ramantani, Louis Maillard, and Laurent Koessler. "Correlation of invasive EEG and scalp EEG." In: *Seizure* 41 (2016), pp. 196–200.

[Rok10] Lior Rokach. "Ensemble-based classifiers." In: *Artificial intelligence review* 33 (2010), pp. 1–39.

[Ron+12] Emanuele Rondonotti et al. "Can we improve the detection rate and inter-observer agreement in capsule endoscopy?" In: *Digestive and Liver Disease* 44.12 (2012), pp. 1006–1011.

[Ruf+19] Lukas Ruff et al. "Deep semi-supervised anomaly detection." In: *arXiv preprint arXiv:1906.02694* (2019).

[Rus+21] Furqan Rustam et al. "Wireless capsule endoscopy bleeding images classification using CNN based model." In: *IEEE access* 9 (2021), pp. 33675–33688.

[Sá+23] Daniel GP de Sá et al. "Abnormality Detection in Wireless Capsule Endoscopy Images Using Deep Features." In: *International Conference on Wireless Mobile Communication and Healthcare*. Springer. 2023, pp. 173–184.

[Sah+22] AWYRC Sahafi et al. "Edge artificial intelligence wireless video capsule endoscopy." In: *Scientific reports* 12.1 (2022), p. 13723.

[Sai+20] Hiroaki Saito et al. "Automatic detection and classification of protruding lesions in wireless capsule endoscopy images based on a deep convolutional neural network." In: *Gastrointestinal endoscopy* 92.1 (2020), pp. 144–151.

[San+18a] Mark Sandler et al. "Mobilenetv2: Inverted residuals and linear bottlenecks." In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 4510–4520.

[San+18b] Shibani Santurkar et al. "How does batch normalization help optimization?" In: *Advances in neural information processing systems* 31 (2018).

[SC13] Saeid Sanei and Jonathon A Chambers. *EEG signal processing*. John Wiley & Sons, 2013.

[Sch+23] Andreas Schulze-Bonhage et al. "Focal cortex stimulation with a novel implantable device and antiseizure outcomes in 2 prospective multicenter single-arm trials." In: *JAMA neurology* 80.6 (2023), pp. 588–596.

[Sch+25] Andreas Schulze-Bonhage et al. "Two-year outcomes of epicranial focal cortex stimulation in pharmacoresistant focal epilepsy." In: *Epilepsia* (2025).

[Sha+20] Karishma Sharma et al. "Noiserank: Unsupervised label noise reduction with dependence models." In: *European conference on computer vision*. Springer. 2020, pp. 737–753.

[She+20] Sheng Shen et al. “Deep convolutional neural networks with ensemble learning and transfer learning for capacity estimation of lithium-ion batteries.” In: *Applied Energy* 260 (2020), p. 114296.

[Shi+22] Haoyue Shi et al. “Loss functions for pose guided person image generation.” In: *Pattern Recognition* 122 (2022), p. 108351.

[Sho+21] Afshin Shoeibi et al. “Epileptic seizures detection using deep learning techniques: a review.” In: *International journal of environmental research and public health* 18.11 (2021), p. 5780.

[Sho09] Ali Hossam Shoeb. “Application of machine learning to epileptic seizure onset detection and treatment.” PhD thesis. Massachusetts Institute of Technology, 2009.

[SK95] Peter Sollich and Anders Krogh. “Learning with ensembles: How overfitting can be useful.” In: *Advances in neural information processing systems* 8 (1995).

[SM13] Laurent Sifre and Stéphane Mallat. “Rotation, scaling and deformation invariant scattering for texture discrimination.” In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2013, pp. 1233–1240.

[Sme+21] Pia H Smedsrud et al. “Kvasir-Capsule, a video capsule endoscopy dataset.” In: *Scientific Data* 8.1 (2021), p. 142.

[SMW08] Felice T Sun, Martha J Morrell, and Robert E Wharen. “Responsive cortical stimulation for the treatment of epilepsy.” In: *Neurotherapeutics* 5 (2008), pp. 68–74.

[Söd05] Johannes Söding. “Protein homology detection by HMM–HMM comparison.” In: *Bioinformatics* 21.7 (2005), pp. 951–960.

[Soh+20] Kihyuk Sohn et al. “Fixmatch: Simplifying semi-supervised learning with consistency and confidence.” In: *Advances in neural information processing systems* 33 (2020), pp. 596–608.

[Sri+22] Abhishek Srivastava et al. “Video capsule endoscopy classification using focal modulation guided convolutional neural network.” In: *2022 IEEE 35th International Symposium on Computer-Based Medical Systems (CBMS)*. IEEE. 2022, pp. 323–328.

[SS02] Bernhard Schölkopf and Alexander J Smola. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2002.

[SST16] Ana-Maria Singeap, Carol Stanciu, and Anca Trifan. “Capsule endoscopy: the road ahead.” In: *World journal of gastroenterology* 22.1 (2016), p. 369.

[Sze+17] Christian Szegedy et al. “Inception-v4, inception-resnet and the impact of residual connections on learning.” In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 31. 1. 2017.

[Sze+20] Vivienne Sze et al. *Efficient processing of deep neural networks*. Springer, 2020.

[Tan+23] Suigu Tang et al. “Transformer-based multi-task learning for classification and segmentation of gastrointestinal tract endoscopic images.” In: *Computers in Biology and Medicine* 157 (2023), p. 106723.

[Thi+19] Roland D Thijs et al. “Epilepsy in adults.” In: *The lancet* 393.10172 (2019), pp. 689–701.

[TNR00] Jonathan Ying Fai Tong, David Nagle, and Rob A Rutenbar. “Reducing power by optimizing the necessary precision/range of floating-point arithmetic.” In: *IEEE Transactions on Very Large Scale Integration (VLSI) Systems* 8.3 (2000), pp. 273–286.

[Tru+18] Nhan Duy Truong et al. “Integer convolutional neural network for seizure detection.” In: *IEEE Journal on Emerging and Selected Topics in Circuits and Systems* 8.4 (2018), pp. 849–857.

[UHA+18] Ihsan Ullah, Muhammad Hussain, Hatim Aboalsamh, et al. "An automated system for epilepsy detection using EEG brain signals based on deep learning approach." In: *Expert Systems with Applications* 107 (2018), pp. 61–71.

[VC17] Juan Vanerio and Pedro Casas. "Ensemble-learning approaches for network security and anomaly detection." In: *Proceedings of the workshop on big data analytics and machine learning for data communication networks*. 2017, pp. 1–6.

[Vit67] Andrew Viterbi. "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm." In: *IEEE transactions on Information Theory* 13.2 (1967), pp. 260–269.

[Wan+04] Zhou Wang et al. "Image quality assessment: from error visibility to structural similarity." In: *IEEE transactions on image processing* 13.4 (2004), pp. 600–612.

[Wan+19] Yisen Wang et al. "Symmetric cross entropy for robust learning with noisy labels." In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2019, pp. 322–330.

[Wan+21] Min Wang et al. "Multipoint simultaneous tracking of wireless capsule endoscope using magnetic sensor array." In: *IEEE transactions on instrumentation and measurement* 70 (2021), pp. 1–10.

[Wei+18] Xiaoyan Wei et al. "Automatic seizure detection using three-dimensional CNN based on multi-channel EEG." In: *BMC medical informatics and decision making* 18 (2018), pp. 71–80.

[Wer+23] Julia Werner et al. "Precise localization within the gi tract by combining classification of cnns and time-series analysis of hmms." In: *International Workshop on Machine Learning in Medical Imaging*. Springer. 2023, pp. 174–183.

[Wer+24] Julia Werner et al. "Energy-Efficient Seizure Detection Suitable for Low-Power Applications." In: *2024 International Joint Conference on Neural Networks (IJCNN)*. IEEE. 2024, pp. 1–8.

[Wer+25a] Julia Werner et al. "Enhanced anomaly detection for capsule endoscopy using ensemble learning strategies." In: *2025 47th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE. 2025, pp. 1–7.

[Wer+25b] Julia Werner et al. "Seeing More with Less: Video Capsule Endoscopy with Multi-task Learning." In: *International Workshop on Applications of Medical AI*. Springer. 2025, pp. 12–21.

[WHO] WHO. *Epilepsy: a public health imperative*. URL: https://www.who.int/publications/i/item/epilepsy-a-public-health-imperative (visited on 02/28/2025).

[WSB03] Zhou Wang, Eero P Simoncelli, and Alan C Bovik. "Multiscale structural similarity for image quality assessment." In: *The thrity-seventh asilomar conference on signals, systems & computers, 2003*. Vol. 2. Ieee. 2003, pp. 1398–1402.

[Wu+10] Hao Wu et al. "Redefining CpG islands using hidden Markov models." In: *Biostatistics* 11.3 (2010), pp. 499–514.

[Xia+15] Tong Xiao et al. "Learning from massive noisy labeled data for image classification." In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 2691–2699.

[Xu+24] Ting Xu et al. "A Multi-task Neural Network for Image Recognition in Magnetically Controlled Capsule Endoscopy." In: *Digestive Diseases and Sciences* (2024), pp. 1–9.

[Ye+12] Yunxing Ye et al. “Accuracy of RSS-based RF localization in multi-capsule endoscopy.” In: *International Journal of Wireless Information Networks* 19 (2012), pp. 229–238.

[YS13] Sehyuk Yim and Metin Sitti. “3-D localization method for a magnetically actuated soft capsule endoscope and its applications.” In: *IEEE Transactions on Robotics* 29.5 (2013), pp. 1139–1151.

[Yu+10] Lean Yu et al. “Support vector machine based multiagent ensemble learning for credit risk evaluation.” In: *Expert Systems with Applications* 37.2 (2010), pp. 1351–1360.

[Yu+21] Xiaoyuan Yu et al. “Multi-task model for esophageal lesion analysis using endoscopic images: classification with image retrieval and segmentation with attention.” In: *Sensors* 22.1 (2021), p. 283.

[Yu02] Marcia Yu. “M2A™ capsule endoscopy: a breakthrough diagnostic tool for small intestine imaging.” In: *Gastroenterology Nursing* 25.1 (2002), pp. 24–27.

[ZBP14] Mingda Zhou, Guanqun Bao, and Kaveh Pahlavan. “Measurement of motion detection of wireless capsule endoscope inside large intestine.” In: *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE. 2014, pp. 5591–5594.

[Zha+19] Chuxu Zhang et al. “A deep neural network for unsupervised anomaly detection and diagnosis in multivariate time series data.” In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 33. 01. 2019, pp. 1409–1416.

[Zho+18] Mengni Zhou et al. “Epileptic seizure detection based on EEG signals and CNN.” In: *Frontiers in neuroinformatics* 12 (2018), p. 95.

[ZM12] Cha Zhang and Yunqian Ma. *Ensemble machine learning*. Vol. 144. Springer, 2012.

[Zon+18] Bo Zong et al. “Deep autoencoding gaussian mixture model for unsupervised anomaly detection.” In: *International conference on learning representations*. 2018.

[Zou+15] Yuexian Zou et al. “Classifying digestive organs in wireless capsule endoscopy images based on deep convolutional neural network.” In: *2015 IEEE International Conference on Digital Signal Processing (DSP)*. IEEE. 2015, pp. 1274–1278.

[ZP17] Chong Zhou and Randy C Paffenroth. “Anomaly detection with robust deep autoencoders.” In: *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*. 2017, pp. 665–674.

[ZS18] Zhilu Zhang and Mert Sabuncu. “Generalized cross entropy loss for training deep neural networks with noisy labels.” In: *Advances in neural information processing systems* 31 (2018).

[Zwi+19] Lilli L Zwinger et al. “CapsoCam SV-1 versus PillCam SB 3 in the detection of obscure gastrointestinal bleeding: results of a prospective randomized comparative multicenter study.” In: *Journal of clinical gastroenterology* 53.3 (2019), e101–e106.

Appendix A

Appendix - Publications

The publications from the author of this dissertation are listed below in the same order as discussed in this thesis. Subsequently, all publications are attached in full length.

Publication A1

Title:	Energy-Efficient Seizure Detection Suitable for Low-Power Applications
Authors:	Julia Werner, Bhavya Kohli, Paul Palomero Bernardo, Christoph Gerum, Oliver Bringmann
Conference:	IEEE International Joint Conference on Neural Networks (IJCNN), 2024
Copyright:	©2024 IEEE, the revised version is available at 10.1109/IJCNN60899.2024.10650710.

Publication A2

Title:	Precise Localization within the GI Tract by Combining Classification of CNNs and Time-Series Analysis of HMMs
Authors:	Julia Werner, Christoph Gerum, Moritz Reiber, Jörg Nick, Oliver Bringmann
Conference:	Medical Image Computing and Computer Assisted Intervention (MICCAI) - International Workshop on Machine Learning in Medical Imaging (MLMI), 2024. Reproduced with permission from Springer Nature.
Copyright:	©2024 Springer Nature Switzerland AG, the revised version is available at https://doi.org/10.1007/978-3-031-45676-3_18.

Publication A3

Title:	Enhanced Anomaly Detection for Capsule Endoscopy Using Ensemble Learning Strategies
Authors:	Julia Werner, Christoph Gerum, Jörg Nick, Maxime Le Floch, Franz Brinkmann, Jochen Hampe, Oliver Bringmann
Conference:	47th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 2025
Copyright:	©2025 IEEE, the revised version is available at 10.1109/EMBC58623.2025.11253055

Publication A4

Title: Seeing More with Less: Video Capsule Endoscopy with Multi-Task Learning

Authors: Julia Werner, Oliver Bause, Julius Oexle, Maxime Le Floch, Franz Brinkmann, Jochen Hampe, Oliver Bringmann

Conference: The 4th Workshop on Applications of Medical AI (AMAI) at MICCAI, 2025. Reproduced with permission from Springer Nature.

Publication A5

Title: Reliable Mislabel Detection for Video Capsule Endoscopy Data

Authors: Julia Werner, Julius Oexle, Oliver Bause, Maxime Le Floch, Franz Brinkmann, Hannah Tolle, Jochen Hampe, Oliver Bringmann

Conference: 48th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 2026

Publication A6

Title: Smart Video Capsule Endoscopy: Raw Image-Based Localization for Enhanced GI Tract Investigation

Authors: Oliver Bause*, Julia Werner*, Paul Palomero Bernardo, Oliver Bringmann (*Equal Contribution)

Conference: 32nd International Conference on Neural Information Processing (ICONIP), 2025. Reproduced with permission from Springer Nature.

Energy-Efficient Seizure Detection Suitable for low-power Applications

Julia Werner*
Department of Computer Science
University of Tübingen
Tübingen, Germany
julia-helga.werner@uni-tuebingen.de
*Corresponding author

Bhavya Kohli
Department of Electrical Engineering
Indian Institute of Technology
Bombay, India
bhavyakohli@iitb.ac.in

Paul Palomero Bernardo
Department of Computer Science
University of Tübingen
Tübingen, Germany
paul.palomero-bernardo@uni-tuebingen.de

Christoph Gerum
Department of Computer Science
University of Tübingen
Tübingen, Germany
christoph.gerum@uni-tuebingen.de

Oliver Bringmann
Department of Computer Science
University of Tübingen
Tübingen, Germany
oliver.bringmann@uni-tuebingen.de

Abstract—Epilepsy is the most common, chronic, neurological disease worldwide and is typically accompanied by reoccurring seizures. Neuro implants can be used for effective treatment by suppressing an upcoming seizure upon detection. Due to the restricted size and limited battery lifetime of those medical devices, the employed approach also needs to be limited in size and have low energy requirements. We present an energy-efficient seizure detection approach involving a TC-ResNet and time-series analysis which is suitable for low-power edge devices. The presented approach allows for accurate seizure detection without preceding feature extraction while considering the stringent hardware requirements of neural implants. The approach is validated using the CHB-MIT Scalp EEG Database with a 32-bit floating point model and a hardware suitable 4-bit fixed point model. The presented method achieves an accuracy of 95.28 %, a sensitivity of 92.34 % and an AUC score of 0.9384 on this dataset with 4-bit fixed point representation. Furthermore, the power consumption of the model is measured with the low-power AI accelerator UltraTrail, which only requires 495 nW on average. Due to this low-power consumption this classification approach is suitable for real-time seizure detection on low-power wearable devices such as neural implants.

Index Terms—Seizure Detection, Time-Series Analysis, CNNs

I. Introduction

Approximately 65-70 million people are affected by epilepsy, one of the most common chronic neurological disease globally [18] [26] [27]. It is characterized by reoccurring seizures which result from an abnormal increased electric activity in the brain. They can lead to short-term mental absence, unconsciousness and convulsions [10] [11]. Reoccurring epileptic seizures not only diminish the overall quality of life of patients but are furthermore a severe safety hazard which can potentially cause severe accidents, in the worst case life-threatening. One common treatment option are anti-epileptic drugs (AED); however, for approximately 30 % [24] of adult patients these do not show any effect or are accompanied by non-bearable side-effects.

For this group of individuals, other approaches need to be explored. One promising alternative is electric brain stimulation, which can be conducted through peripheral nerve stimulation, spinal cord stimulation or deep brain stimulation [12] [21]. Devices for electric brain stimulation can either perform continuous stimulation or responsive stimulation upon seizure detection, and in both cases patients do not recognize the stimulation [3] [12]. For example, one such device for neuromodulation on the market is the Responsive Neurostimulation System (RNS) of Neuropace [24] which is a neural implant with four electrodes that is implanted within the brain at a seizure onset region and recognizes as well as interrupts the beginning of a focal seizure. However, seizure devices are generally limited in size and therefore the available energy is limited. Each addition to the system requires its share of energy. If an algorithm is employed directly on the device, the total amount of area, number of computations and amount of energy has to be restricted as much as possible to limit the total energy consumption and prolong the battery lifetime.

This paper presents an energy-efficient approach to accurately determine anomalies within recorded Electroencephalography (EEG) data which are classified as seizures. By combining a light-weight neural network without preceding feature extraction but with subsequent time-series analysis methods, accurate anomaly/seizure detection is ensured while simultaneously providing a model with low complexity making it suitable for low-power applications. The model is validated using the CHB-MIT Scalp EEG

This work has been partly funded by the German Federal Ministry of Education and Research (BMBF) in the project MEDGE (16ME0530).

Database collected at the Children's Hospital Boston [23] [19]. This enables subsequent deployment to low-power hardware accelerators for real-time seizure detection which is demonstrated on the UltraTrail hardware architecture [4].

A. Related work

Shoeb and Guttag published the CHB-MIT dataset in 2009 and achieved a sensitivity of 96 % of 173 test seizures with a mean latency of 4.6 s [23] while using Support Vector Machines (SVMs). Subsequently, others have used this dataset for further experiments but only few simulated the models on hardware. Hügle et al. [13] presented SeizureNet which was executed on hardware with a power consumption of 850 µW. Bahr et al. [2] tested the classification of a CNN and a low-power microprocessor with this dataset. They achieved a sensitivity of 85 % with the CNN on a microcontroller and an average power of 140 µW. Zanetti et al. [31] used random forest models on an ARM cortex-M4 microcontroller and achieved a sensitivity of 96.6 %, with a battery lifetime of 40.87 hours and 7.34 mA. Manzouri et al. tested a LSTM, CNN and a random forest classifier an ultra low-power microcontroller [17] with a dataset of the Epilepsy Center Freiburg resulting in 7 µW for the CNN approach.

For this application, some research has been conducted in combination with time-series analysis. For example, Craley et al. [7] presented a coupled Hidden Markov Model for seizure detection and Dash and Kolekar [8] also performed seizure detection involving HMMs and Viterbi decoding, achieving an accuracy of 96.87 %. However, this requires preceding feature extraction and furthermore, no hardware simulation was conducted and no energy estimation was provided in their experiments. Furthermore, there have been some usages of smoothing data with a moving average approach in the context of seizure detection in the past [1] [22] [25] [30]. Rana et al. [22] have employed the moving average method in the context of seizure detection. Although it was not used in combination with a CNN; but functioned as a tool for establishing a seizure detection threshold in their proposed method. Yu et al. [30] combine a CNN for feature extraction with principal components analysis, Bayesian linear discriminant analysis and moving average filtering. Temko et al. [25] use extracted feature-vectors with support-vector machines (SVMs) in combination with moving average smoothing.

We propose an energy-efficient and low-complex method which combines a very light-weight CNN with simple time-series analysis without the need of preceding feature extraction that is suitable for low-power hardware architectures. We validate this using the CHB-MIT dataset and the UltraTrail hardware accelerator.

II. Methodology

An overview of the complete classification approach is shown in Fig. 1. First, the TC-ResNet4 is trained on preprocessed EEG data but without preceding feature extraction. After completion of the retraining, the model is used to classify incoming EEG data; subsequently, time-series analysis is employed either in the form of Viterbi decoding and a HMM or as a simple smoothing approach with simple moving average or exponentially moving average. Finally, for each EEG fragment, the classification approach returns either a 0 for non-ictal recordings or 1, if a seizure was detected.

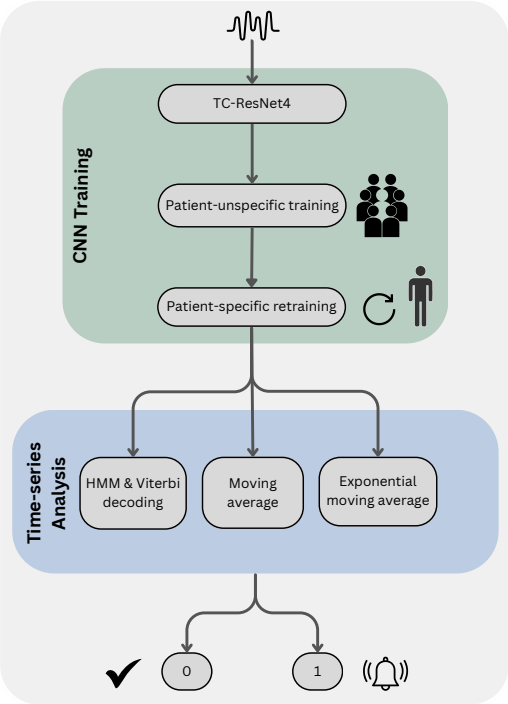


Fig. 1: Overview of the classification approach.

Dataset description

For all experiments, the CHB-MIT dataset [19] [23] was used, which contains the EEG data from 24 patients stored in the European Data Format (as .edf files). In each patients' respective directory, there are multiple such .edf files containing EEG data for up to 1 hour, read from different channels. The file annotations are provided separately, and the annotations contain the time-stamps for the start and end of a seizure.

A. Preprocessing

To account for individual differences in EEG data, the main objective is to perform patient-specific retraining using a pre-trained base model and patient specific data. Thus, the data was split in the following way for each patient:

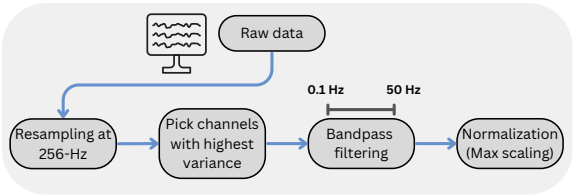


Fig. 2: Preprocessing pipeline.

1) The data was originally sampled at 256 Hz, any files with a different sampling rate were resampled to 256 Hz.
2) Not all channels found in the dataset are present in each individual patient data. Therefore, only common channels across all patients were chosen to ensure uniform representation. From those, the number of channels was further reduced to 16 based on the channels with the highest variance in the ictal data as suggested by [9] to exploit hardware advantages.
3) To remove DC components and the noisy component from the EEG measurement device, a band-pass filter was applied between 0.1 Hz and 50 Hz. Following the filter, the data was split into fragments of length 0.5 s, or correspondingly, fragments of 128 samples.
4) The complete data for a patient was collected and labelled according to the extracted annotations, giving a large array of shape $(N \times 128 \times 16)$.
5) Due to the scarcity of ictal data compared to non-ictal data, to maintain a reasonable ratio of positives and negatives, the non-ictal data was sampled such that for the dev and retrain set the final ratio of ictal to non-ictal data was 3. However, this ratio was not enforced in the test set in order to test on a realistic proportion for each class.
6) For the test set, for each patient one complete edf file consisting of an EEG recording with at least one seizure was collected.
7) The remaining data was split into two stratified (to maintain the specified class ratio) subsets in a $40:60$ ratio. The two subsets are described below:
   a) dev (40 %): Data in this subset is first collected from all patients, and then further split into a $80:20$ ratio (training and validation) for patient-unspecific training.
   b) retrain (60 %): Data in this subset is individually split into a $80:20$ ratio (training and validation), and is used for patient-specific retraining.
8) The data in the test set is used purely for testing. The collected test data from all patients is used during the patient-unspecific training phase, and the individual patient data is used during patient-specific retraining.
9) For normalization, the absolute maximum value of the train set, validation set and test set was identified and the maximum value among those was used for normalizing each subset before training the neural network.

Fig. 2 illustrates the basic preprocessing pipeline for a given patient. The final class distribution for the complete train and test set, and average class distribution of retrain set and test set for all patients is reported in Table I. This also includes the data samples of the 2 randomly selected patients, that is needed for training the time-series methods and thus excluded from final testing.

TABLE I: Class distribution of the (half second fragments) dataset used.
(* average data per patient)

Subset	Negative samples	Positive Samples	Total Samples
dev	16,017	5,339	21,356
retrain	24,111	8,037	32,148
test	258,542	4,796	263,338
retrain*	1,005	335	1,334
test*	10,773	200	10,973

B. Classification approach

Classification with a CNN

Anomaly detection of EEG data is particularly useful in conjunction with a neural implant, so that it can act as a seizure detection device. However, the limited size of such implants restricts also the model size, e.g. in terms of the number of parameters. Thus, it is crucial to find a model which is suitable for low-power devices and simultaneously exhibits a sufficient accuracy and sensitivity. By combining temporal convolutional neural networks (TCNs) with residual networks (ResNets) the TC-ResNets [6] are generated. TC-ResNets are light-weight 1D-convolutional neural networks with relatively few parameters which were successfully used for sensor-signal processing tasks [6]. For this approach, it is beneficial that CNNs do not require a preceding feature extraction. Since TC-ResNets are characterized by a low complexity and additionally have been successfully applied to other problems [32], these CNNs were used as a base model. During the training process, it became apparent that a TC-ResNet4 provides the best trade-off between accuracy and complexity and was therefore used for further experiments without preceding feature extraction. The TC-ResNet4 archictecture [6] is presented in Table II.

The training was partitioned into two phases, considering at each step the suitability to implement this in hardware.

(1) The first phase of training involves using the collected data from all patients to train a patient-unspecific base model. Training was performed for 40 epochs with a batch size of 128 and the AdamW optimizer with a

TABLE II: TC-ResNet4 architecture

Layer Type	Output	Params	MAC ops
Input	16×128	0	0
Conv1d	16×63	768	48K
Conv1d	24×32	3,456	111K
BatchNorm1d	24×32	0	0
Hardtanh	24×32	0	0
Conv1d	24×32	5,184	166K
BatchNorm1d	24×32	0	0
Conv1d	24×32	384	12.3K
BatchNorm1d	24×32	0	0
ReLU	24×32	0	0
Hardtanh	24×32	0	0
TCResidualBlock	24×32	0	0
GlobalAveragePooling1d	24×1	0	0
Dropout	24	0	0
Linear	2	48	48

learning rate of 0.001. For computing the loss, the different class frequencies were considered to calculate the weights for the loss function. Furthermore, we employed threshold moving [20] [33] to modify the default threshold of 0.5 when classifying the data samples. To tune the threshold, we used the evaluations of the TC-ResNet4 on the train set by weighing the output probability of detecting a seizure with a weight $w \in \{1, 2, 3, 4, 5\}$. The smallest weight that achieved a sensitivity larger than 0.9 in our experiments was $w = 2$, which was therefore the hyperparameter of choice. Hence, this was employed for all following experiments, resulting in a slightly higher sensitivity at the cost of a marginally decreased specificity.

(2) For the patient-specific retraining in the second phase, the base model is loaded and then retrained on data from individual patients for additional 10 epochs with the same configurations as the base model, except for a batch size of 8. After the patient-specific TC-ResNet4 is fully trained, three different approaches of time-series analysis were used to enhance the given predictions of the TC-ResNet4. The experiments were conducted once with 32-bit floating point computation as well as with 4-bit fixed point representation to meet the hardware requirements. As a final step, time-series analysis is performed for each patient.

Classification with time-series analysis

1) Simple Moving Average: The main reason to use subsequent time-series analysis was to improve the classification result of the CNN by smoothing the series of data points with a preferably computationally cheap method. One simple yet popular approach for this is applying a simple moving average (SMA) method [5]. For this method, for a series of data points X, for a window size w the mean is computed as follows:

$$SMA = \frac{1}{w} \sum_{n-w+1}^{n} X_i.$$

In this setting, the SMA was applied with a window size $w = 5$ to the final evaluations of the TC-ResNet4, the probabilities are finally converted to binary labels.

2) Exponentially Moving Average: The exponentially moving average [14] (EWMA) provides exponentially weighted moving averages to put more emphasis on the latest observations of the neural network and was additionally applied. To determine the threshold for both methods, two randomly selected patients were used to obtain the best threshold based on a grid search and were further excluded from testing in all experiments.

3) Hidden Markov Model: Time-series analysis in the form of a Hidden Markov Model (HMM) and Viterbi decoding provides a hyperparameter free method, which therefore does not require many samples for training. This is a major advantage compared to other models such as LSTMs and the main reason for choosing this method. With this approach, the predicted labels of the TC-ResNet4 are passed on as integers to the HMM and subsequent Viterbi decoding returns the most likely sequence of hidden states based on the provided observations given by the TC-ResNet, similarly as employed in [29]. Due to the backtracking approach of the Viterbi algorithm, this detection method is accompanied by a delay which describes the time between the first occurrence of a seizure and the actual detection by the algorithm and mainly depends on the chosen window size for the included data. For the presented approach, the same window size $w = 5$ was chosen as for the other methods, which results in a maximal delay of 2.5 s for 0.5 s fragments introduced by the Viterbi decoding. The transition and emission probabilities were acquired solely based on statistical measurements. To compute the transition probabilities, after the EEG data was split into 0.5 s fragments, the number of transitions from each class to the other class were counted for the data from the train and retrain set [15]. The proportions can then directly be encoded in the transition probabilities of the HMM. The confusion matrix of the CNN classifying the training data in the last epoch of training naturally encodes the emission probabilities for this setting, thus the emission probabilities of the HMM were computed from this confusion matrix.

Metrics/Evaluation

The accuracy, sensitivity, specificity as well as the false-positive rate (FPR) were computed based on the confusion matrix of the final model. Furthermore, the receiver operating characteristic curve area under the curve (AUC) score was computed with the scikit-learn metrics library.

C. Inference

To demonstrate the low-power consumption of this light-weight neural network for future applications, a low-power real-time AI accelerator for sensor-signal processing seems to be the most suitable hardware architecture. The AI accelerator UltraTrail, which previously has been successfully used for TC-ResNets on keyword spotting tasks [4], has been employed for the described neural network. One benefit of this hardware architecture is that it is optimized for TC-ResNet topologies, involving a configurable array of processing elemens and a distributed memory system with dynamic content re-allocation [4]. The model was trained with quantization-aware training and finally, the PyTorch code was converted to C-code. For hardware execution, 4-bit quantization was used for computing the weights, bias and features in fixed-point representation.

III. Results

As described in Section II-B, the TC-ResNet4 was first regularly trained for 40 epochs on data from all patients and then retrained on the data from each patient for 10 epochs with a batch size of 8 with 0.5 s fragments. First, the results for the non-quantized TC-ResNet4 with 32-bit floating point number computation is shown in Table III. It is demonstrated that each of the three time-series methods provide proficient classification results. The HMM provides the highest AUC score of 0.9530, while the exponentially weighted moving average method provides the highest sensitivity with 93.71 %. However, in terms of accuracy, specificity and AUC score, the exponential moving average method, which gives more weight to the latest observations, is characterized by a slightly poorer performance. One possible reason for this is that the window size of 5 might be too restrictive for this method. However, it is important to note, that although models are typically trained with 32-bit floating-point number computation, which allows higher precision, this is not suitable for low-power hardware applications. To provide a hardware-suitable method, we explored how different word widths of the weights of the TC-ResNet affect the overall classification ability on these different approaches.

TABLE III: Results of the non-quantized TC-ResNet combined with different subsequent time-series analysis methods.

	TC-ResNet4 (32-bit floating-point)		
	Moving Average	Exp. Moving Average	HMM
Accuracy [%]	96.05	94.16	97.84
Sensitivity [%]	92.04	93.71	92.67
Specificity [%]	96.12	94.17	97.93
FPR	0.0388	0.0583	0.0207
AUC score	0.9408	0.9394	0.9530

The AUC score was computed for 1) the TC-ResNet4 base model (without retraining), 2) the same model with additional retraining, and 3) the retrained model with subsequent time-series analysis using Viterbi decoding after quantization-aware training was performed. The results are presented in Fig. 3 for the word widths $\{2, 4, 6, 8, 10\}$. This demonstrates on the one hand that 4 bits are sufficient to achieve good results and concurrently that the base model on its own is capable of good classification. Nevertheless, incorporating additional patient-specific retraining and adding time-series analysis increases the precision even further.

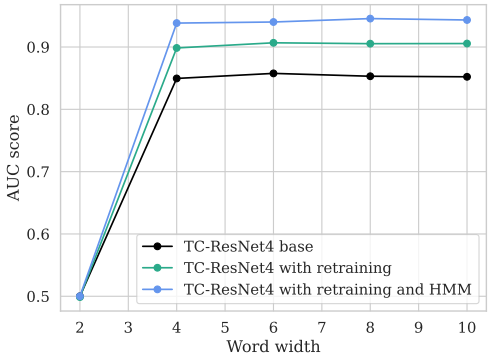


Fig. 3: AUC scores computed for the base model, the retrained model and the retrained model with subsequent Viterbi decoding while the weights of the model were quantized to different bits.

Subsequently, for the following experiments, quantization-aware training was performed for the base model as well as for the patient-specific model with a word width of 4 bit allocated for each weight. Table IV presents the results of the trained and quantized TC-ResNet4 with the subsequent time-series analysis approaches as described in Section II. All methods perform slightly weaker than if 32 bits were used. This is expected due to a higher precision when computing with a larger number of bits. Nevertheless, incorporating quantization in the training process seems to compensate this and considering the hardware suitability, using 4 instead of 32 bits is a more reasonable approach for this setting.

For each metric, the HMM provides the best results compared to the moving average approaches with 4-bit fixed point representation. However, all three time-series analysis methods show similarly solid classification abilities. Additionally, for each patient the AUC scores were computed and plotted for each of the time-series methods as shown in Fig. 4. The main outlier was the patient with the ID 16, who also sticks out in the remaining results section.

Furthermore, the sensitivity, specificity and accuracy were plotted for each patient for the classification approach

TABLE IV: Results of the quantized TC-ResNet4 combined with different subsequent time-series analysis methods.

	Quantized TC-ResNet4 (4-bit fixed-point)		
	Moving Average	Exp. Moving Average	HMM
Accuracy [%]	92.79	94.69	95.28
Sensitivity [%]	92.15	90.20	92.34
Specificity [%]	92.80	94.77	95.34
FPR	0.0720	0.0523	0.0466
AUC score	0.9247	0.9248	0.9384

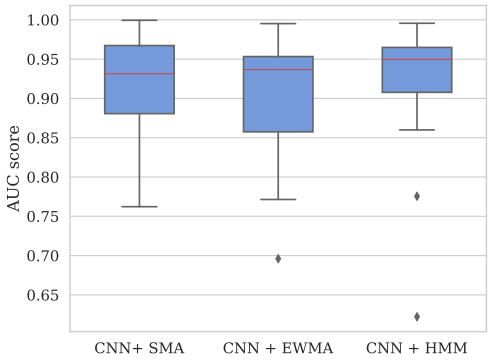


Fig. 4: AUC scores computed for each patient and for each of the three time-series analysis method: SMA, EWMA and HMM.

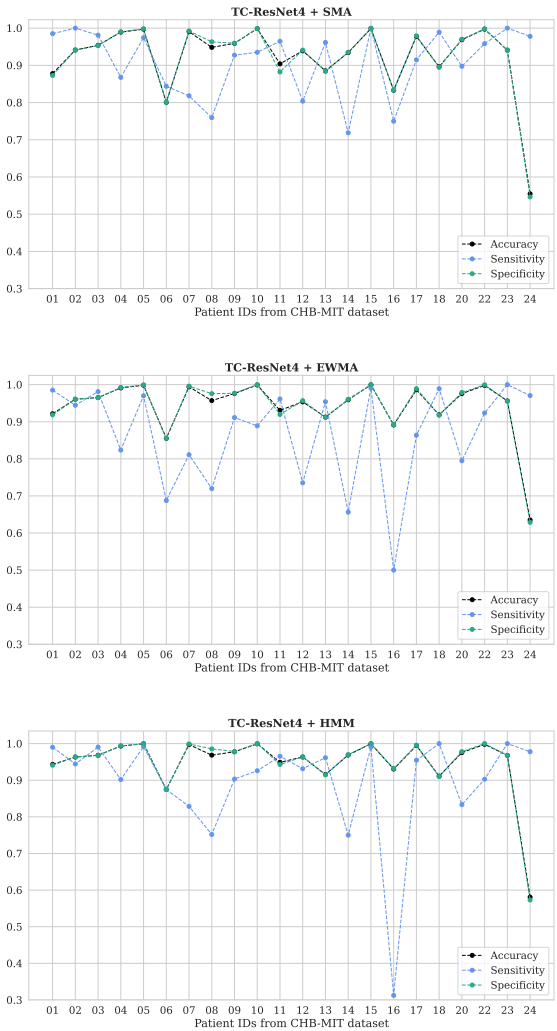


Fig. 5: Metrics shown for each patient and for each of the three approaches in combination with the 4-bit quantized TC-ResNet4.

involving the CNN and the different time-series methods as depicted in Fig. 5. In particular, for the SMA and the EWMA, the sensitivity values are the lowest compared to the other metrics. Across all three time-series methods, the patient with the ID 16 exhibits inferior results compared to the others. Interestingly, this patient also has the lowest number of seizure samples ($n = 57$) in the retraining set, indicating that these are not sufficient samples to retrain the model perfectly. We further observe that the HMM in combination with Viterbi decoding performs slightly better across all patients (except for patient 16) than the moving average methods. The mean detection delay of a seizure for the HMM approach was 4.41 s, including the delay introduced by the Viterbi decoding.

We validated that a patient-specific model achieves superior results compared to a patient-unspecific model. If this is realized in the future, it is important to note that the performance of the model improves with a larger amount of acquired data for each patient. The presented experiments suggest that it is beneficial to acquire some seizure data for each patient. On average, 335 half-seconds of seizure data were used for retraining per patient in the presented experiments. However, if acquiring this data is not feasible, our results further indicate that simply using a trained base model in combination with the time-series analysis will provide sufficient results for the majority of patients as well.

A. Inference

The results of the Viterbi decoding for the described application can be directly implemented in a look up table (LUT) with a size of k^{n+1}, for k states and a window size of n. There are 2^n possible combinations for the n predictions of the CNN, which the HMM receives. For each sequence of predictions, a computed prediction of the HMM can be generated based on the Viterbi algorithm. The result for

this can be stored as a $(n+1)^{\text{th}}$ entry in the LUT. For our application we have two possible states (ictal, non-ictal) and a window size of 5. Thus, the total size of this LUT would be 64. The simple moving average method with a window size of 5 requires the additional storage of 6 integer values. Both methods only require a small amount of additional storage and basically no increased power consumption. Therefore, they are negligible and have not been included in the energy estimates.

The employed UltraTrail architecture is shown in Figure 6. This architecture consists of a control unit, three feature memories (FMEM), one interconnect unit, a local memory (LMEM), an output processing unit (OPU), the bias memory (BMEM) and a weight memory (WMEM) as published by Palomero et al. [4]. The FMEM stores the input and output features as well as the residual paths, while the WMEM and BMEM store all DNN parameters on chip. The multiply-and-accumulate units of the MAC Array compute the matrix multiplications of the network. The LMEM then stores the partial sums for the accumulation inside the MAC Array. The OPU is for example used for computing the activations, bias or pooling functions. In our experiments, the architecture is adapted to our specific use case and to the presented TC-ResNet4 by modifying the MAC Array to a 4×4 instead of a 8×8 MAC Array and adjusting the memory sizes accordingly, as shown in Figure 6.

For evaluation, the design was synthesized in GlobalFoundries 22FDX® 22 nm FD-SOI technology using Cadence Genus. The power estimation was conducted for typical conditions (25 °C, 0.8 V) using Cadence Joules. Executing the quantized TC-ResNet4 on UltraTrail [4] with 10 inferences per second and a clock frequency of 250 kHz results in a total power consumption of 495 nW on average. The clock frequency can be adjusted depending on the real-time requirements of the use case. Executing the network requires 80.626 ms, afterwards the accelerator switches into a power gating state until the next inference starts. The total required area is 26 319.79 µm^2. Fig. 7 depicts the area cost and the power consumption for each of the elements in detail. The memory wrapper consists of the FMEM, the WMEM, and the BMEM. In total, the CNN consists of 9840 parameters and requires 337968 MAC operations.

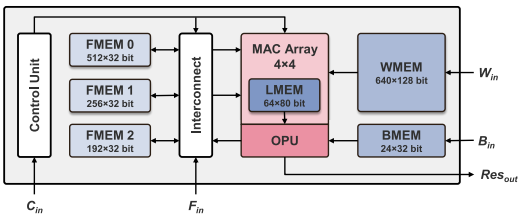


Fig. 6: UltraTrail architecture [4] with a 4×4 MAC Array.

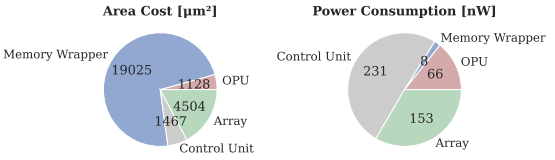


Fig. 7: Area costs and power consumption for different hardware components of the UltraTrail accelerator.

The measured power consumption in comparison to results from other work is listed in Table V. However, comparing those power values is challenging, due to missing information or different types of measurement. Hügle et al. [13] used a MSP430FR599 microcontroller with a 32-bit hardware multiplier achieving 850 µW for one inference. On UltraTrail, we need notably less power with only 0.495 µW at 10 Hz, including the idle periods when the system is waiting for the next computation. Furthermore, our approach provides also advantages if used with other architectures as an alternative to UltraTrail. For example, our employed 4-bit implementation is beneficial in combination with the Single Instruction Multiple Data (SIMD) technique. This theoretically allows the processing of 8×4-bit operations in parallel on a 32-bit hardware multiplier using a single 32-bit register, which potentially leads to an energy reduction by a factor of 4. Bahr et al. [2] use a CNN on a RISC-V based GAP microcontroller, which consumes an average power of 140 µW for one inference on 1 second EEG data. The execution time amounts to 35 ms. Although we have a higher execution time, our power consumption is notably less. Truong et al. [28] compare 32-bit models to a 4-bit IntegerNet in their work, which demands $34-90$ µJ for each inference. However, the specific hardware architecture used for measuring was not specified, which complicates a comparison. We have a total power consumption of 0.495 µW with 10 Hz corresponding to an energy demand of 49.5 nJ for one inference including the idle time. Kiral-Kornek et al. [16] deployed their neural network to an ultra low-power TrueNorth chip and achieved a power consumption of <40 mW. Although they have not provided a concrete execution time, it is evident that our deployment combination is more efficient. Manzouri et al. [17] employed an ultra low-power Apollo 3 Blue microcontroller from Ambiq and had a power consumption of only 0.495 µW for a random forest model and 7.01 µW with a CNN. As they used a classification rate of 1 Hz, this corresponds to 0.495 µJ and 7.01 µJ, respectively. In their work it becomes evident that the random forest model performs inferior in the classification task as compared to the CNN. Nevertheless, our total energy demand including the idle time is notably less with only 0.0495 µJ.

TABLE V: Comparison of results to literature.

Authors	Model	HW-Architecture	Total power
Hügle et al. [13]	SeizureNet	MSP430FR599	850 μW
Bahr et. al [2]	CNN	RISCV based GAP8	140 μW
Truong et. al [28]	IntegerNet	microcontroller	$34 - 90$ μJ*
Kiral-Kornek et al. [16]	DNN	IBM TrueNorth	< 40 mW
Manzouri et al. [17]	CNN	Ambiq Apollo 3 Blue	7 μW
This work	TC-ResNet4	UltraTrail [4]	0.495 μW

*for each classification

IV. Conclusion

Efficient seizure detection is essential for patients suffering from epilepsy who do not respond to drug treatment. In this work, a light-weight 1D-CNN with different time-series analysis techniques was combined, and validated using the CHB-MIT dataset achieving an accuracy of 95.28 %, a sensitivity of 92.34 % and an AUC score of 0.9384 while allocating only 4 bit per weight. The presented approach neglects the necessity of preceding feature extraction. Furthermore, it was demonstrated, that the classification model is suitable for real-time seizure detection by executing it on a low-power hardware architecture with a power consumption of only 495 nW on average for one inference including the remaining idle time with 10 Hz. Especially compared to current baselines for this classification task, the low-power consumption with this approach stands out.

References

[1] Alotaiby, T.N., Alshebeili, S.A., Abd El-Samie, F.E., Alabdulrazak, A., Alkhnaian, E.: Channel selection and seizure detection using a statistical approach. In: 2016 5th international conference on electronic devices, systems and applications (ICEDSA). pp. 1–4. IEEE (2016)

[2] Bahr, A., Schneider, M., Francis, M.A., Lehmann, H.M., Barg, I., Buschhoff, A.S., Wulff, P., Strunskus, T., Faupel, F.: Epileptic seizure detection on an ultra-low-power embedded risc-v processor using a convolutional neural network. Biosensors 11(7), 203 (2021)

[3] Bergey, G.K., Morrell, M.J., Mizrahi, E.M., Goldman, A., King-Stephens, D., Nair, D., Srinivasan, S., Jobst, B., Gross, R.E., Shields, D.C., et al.: Long-term treatment with responsive brain stimulation in adults with refractory partial seizures. Neurology 84(8), 810–817 (2015)

[4] Bernardo, P.P., Gerum, C., Frischknecht, A., Lübeck, K., Bringmann, O.: UltraTrail: A configurable ultralow-power tc-resnet ai accelerator for efficient keyword spotting. IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems 39(11), 4240–4251 (2020)

[5] Box, G.E., Jenkins, G.M., Reinsel, G.C., Ljung, G.M.: Time series analysis: forecasting and control. John Wiley & Sons (2015)

[6] Choi, S., Seo, S., Shin, B., Byun, H., Kersner, M., Kim, B., Kim, D., Ha, S.: Temporal convolution for real-time keyword spotting on mobile devices. arXiv preprint arXiv:1904.03814 (2019)

[7] Craley, J., Johnson, E., Venkataraman, A.: A novel method for epileptic seizure detection using coupled hidden markov models. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part III 11. pp. 482–489. Springer (2018)

[8] Dash, D.P., Kolekar, M.H.: Hidden markov model based epileptic seizure detection using tunable q wavelet transform. Journal of biomedical research 34(3), 170 (2020)

[9] Duun-Henriksen, J., Kjaer, T.W., Madsen, R.E., Remvig, L.S., Thomsen, C.E., Sorensen, H.B.D.: Channel selection for automatic seizure detection. Clinical Neurophysiology 123(1), 84–92 (2012)

[10] Engel, J.: Seizures and epilepsy, vol. 83. Oxford University Press, USA (2013)

[11] Fisher, R.S., Boas, W.V.E., Blume, W., Elger, C., Genton, P., Lee, P., Engel Jr, J.: Epileptic seizures and epilepsy: definitions proposed by the international league against epilepsy (ilae) and the international bureau for epilepsy (ibe). Epilepsia 46(4), 470–472 (2005)

[12] Fisher, R.S., Velasco, A.L.: Electrical brain stimulation for epilepsy. Nature Reviews Neurology 10(5), 261–270 (2014)

[13] Hügle, M., Heller, S., Watter, M., Blum, M., Manzouri, F., Dumpelmann, M., Schulze-Bonhage, A., Woias, P., Boedecker, J.: Early seizure detection with an energy-efficient convolutional neural network on an implantable microcontroller. In: 2018 International Joint Conference on Neural Networks (IJCNN). pp. 1–7. IEEE (2018)

[14] Hunter, J.S.: The exponentially weighted moving average. Journal of quality technology 18(4), 203–210 (1986)

[15] Kacprzyk, J.: Advances in intelligent systems and computing. Springer (2012)

[16] Kiral-Kornek, I., Roy, S., Nurse, E., Mashford, B., Karoly, P., Carroll, T., Payne, D., Saha, S., Baldassano, S., O'Brien, T., et al.: Epileptic seizure prediction using big data and deep learning: toward a mobile system. EBioMedicine 27, 103–111 (2018)

[17] Manzouri, F., Zöllin, M., Schillinger, S., Dümpelmann, M., Mikut, R., Woias, P., Comella, L.M., Schulze-Bonhage, A.: A comparison of energy-efficient seizure detectors for implantable neurostimulation devices. Frontiers in Neurology 12, 703797 (2022)

[18] Moshé, S.L., Perucca, E., Ryvlin, P., Tomson, T.: Epilepsy: new advances. The Lancet 385(9971), 884–898 (2015)

[19] PhysioBank, P.: Physionet: components of a new research resource for complex physiologic signals. Circulation 101(23), e215–e220 (2000)

[20] Provost, F.: Machine learning from imbalanced data sets 101 (2008)

[21] Racine, R.J.: Modification of seizure activity by electrical stimulation: Ii. motor seizure. Electroencephalography and clinical neurophysiology 32(3), 281–294 (1972)

[22] Rana, P., Lipor, J., Lee, H., Van Drongelen, W., Kohrman, M.H., Van Veen, B.: Seizure detection using the phase-slope index and multichannel ecog. IEEE Transactions on Biomedical Engineering 59(4), 1125–1134 (2012)

[23] Shoeb, A.H.: Application of machine learning to epileptic seizure onset detection and treatment. Ph.D. thesis, Massachusetts Institute of Technology (2009)

[24] Skarpaas, T.L., Jarosiewicz, B., Morrell, M.J.: Brain-responsive neurostimulation for epilepsy (rns® system). Epilepsy research 153, 68–70 (2019)

[25] Temko, A., Thomas, E., Marnane, W., Lightbody, G., Boylan, G.: Eeg-based neonatal seizure detection with support vector machines. Clinical Neurophysiology 122(3), 464–473 (2011)

[26] Thijs, R.D., Surges, R., O'Brien, T.J., Sander, J.W.: Epilepsy in adults. The Lancet 393(10172), 689–701 (2019)

[27] Thurman, D.J., Beghi, E., Begley, C.E., Berg, A.T., Buchhalter, J.R., Ding, D., Hesdorffer, D.C., Hauser, W.A., Kazis, L., Kobau, R., et al.: Standards for epidemiologic studies and surveillance of epilepsy. Epilepsia 52, 2–26 (2011)

[28] Truong, N.D., Nguyen, A.D., Kuhlmann, L., Bonyadi, M.R., Yang, J., Ippolito, S., Kavehei, O.: Integer convolutional neural network for seizure detection. IEEE Journal on Emerging and Selected Topics in Circuits and Systems 8(4), 849–857 (2018)

[29] Werner, J., Gerum, C., Reiber, M., Nick, J., Bringmann, O.: Precise localization within the gi tract by combining classification of cnns and time-series analysis of hmms. In: International Workshop on Machine Learning in Medical Imaging. pp. 174–183. Springer (2023)

[30] Yu, Z., Nie, W., Zhou, W., Xu, F., Yuan, S., Leng, Y., Yuan, Q.: Epileptic seizure prediction based on local mean decompo-

sition and deep convolutional neural network. The Journal of Supercomputing 76, 3462–3476 (2020)
[31] Zanetti, R., Aminifar, A., Atienza, D.: Robust epileptic seizure detection on wearable systems with reduced false-alarm rate. In: 2020 42nd annual international conference of the IEEE engineering in medicine & biology society (EMBC). pp. 4248–4251. IEEE (2020)
[32] Zhang, B., Li, W., Li, Q., Zhuang, W., Chu, X., Wang, Y.: Autokws: Keyword spotting with differentiable architecture search. In: ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 2830–2834. IEEE (2021)
[33] Zhou, Z.H., Liu, X.Y.: Training cost-sensitive neural networks with methods addressing the class imbalance problem. IEEE Transactions on knowledge and data engineering 18(1), 63–77 (2005)

Precise localization within the GI tract by combining classification of CNNs and time-series analysis of HMMs.*

Julia Werner[1], Christoph Gerum[1], Moritz Reiber[1], Jörg Nick[2], and Oliver Bringmann[1]

[1] Department of Computer Science, University of Tübingen, Germany.
[2] Department of Mathematics, ETH Zürich, Switzerland.

Abstract. This paper presents a method to efficiently classify the gastroenterologic section of images derived from Video Capsule Endoscopy (VCE) studies by exploring the combination of a Convolutional Neural Network (CNN) for classification with the time-series analysis properties of a Hidden Markov Model (HMM). It is demonstrated that successive time-series analysis identifies and corrects errors in the CNN output. Our approach achieves an accuracy of 98.04% on the Rhode Island (RI) Gastroenterology dataset. This allows for precise localization within the gastrointestinal (GI) tract while requiring only approximately 1M parameters and thus, provides a method suitable for low power devices.

Keywords: Medical Image Analysis · Wireless Capsule Endoscopy · GI Tract Localization.

1 Introduction

The capsule endoscopy is a medical procedure that has been used for investigating the midsection of the GI tract since early 2000 [12, 3]. This minimally invasive method allows to visualize the small intestine, which is in most part not accessible through standard techniques using flexible endoscopes [22]. The procedure starts by swallowing a pill-sized capsule. While it moves through the GI tract by peristalsis, it sends captured images from an integrated camera with either an adaptive or a defined frame rate to an electronic device. The overall aim of this procedure is to detect diseases affecting the small intestine such as tumors and its preliminary stages, angiectasias as well as chronic diseases [22, 17, 24]. Since the esophagus, stomach and colon can be more easily assessed by standard techniques, the small intestine section is of main interest in VCE studies.

All images of the small intestine should be transmitted for further evaluation by medical experts who are qualified to check for anomalies. The frame rate of the most prominent capsules ranges from 1 to 30 frames per second with a varying

* This work has been partly funded by the German Federal Ministry of Education and Research (BMBF) in the project MEDGE (16ME0530)

resolution between 256×256 and 512×512 depending on the platform [22]. For example, the PillCam® SB3 by Medtronic lasts up to 12 hours with an adaptive frame rate of 2 to 6 frames per second [18]. This should ensure passing through the whole GI tract before the energy of the capsule's battery is depleted. However, a capsule can also require more than one day to pass through the whole GI tract leading to an incomplete record of images due to depletion of the capsule's battery after maximal 12 hours. In this procedure, the energy is the bottleneck and small changes of the architecture can increase the overall energy requirement leading to a shorter battery lifetime with the risk of running out of energy without covering the small intestine. However, modifications such as capturing images with a higher resolution might improve the recognition ability of clinicians and thus, it is desirable to increase the limited resolution or add more functions (e.g. zooming in or out, anomaly detection on-site) helping to successfully scan the GI tract for anomalies at the cost of increasing energy demands. The images taken before the small intestine are not of interest but demand their share of energy for capturing and transmitting the images.

This paper presents a method for very accurately determining the location of the capsule by on-site evaluation using a combination of neural network classification and time-series analysis by a HMM. This neglects the necessity to consume electric energy for transmitting images of no interest. If this approach is integrated into the capsule it can perform precise self-localization and the transition from the stomach to the small intestine is verified with high confidence. From this moment onwards, all frames should be send out for further evaluation. A major part of the energy can be saved since the data transmission only starts after the capsule enters the small intestine and therefore can be used for other valuable tasks. For example, the frame rate or resolution could be increased while in the small intestine or additionally, a more complex network for detecting anomalies on-site could be employed.

1.1 Related work

In the field of gastroenterology, there have been different approaches to perform localization of a capsule within the GI tract [16] including but not limited to magnetic tracking [19, 26], video-based [28, 15] and electromagnetic wave techniques [27, 7]. However, Charoen et al [2] were the first to publish a dataset with millions of images classified into the different sections of the GI tract. They achieved an accuracy of 97.1% with an Inception ResNet V2 [23] architecture on the RI dataset and therefore successfully demonstrated precise localization without aiming for an efficient realization on hardware. To the best of our knowledge, there is no superior result than the baseline with this dataset. However, a large network with 56M parameters as the Inception ResNet V2 is not suitable for low-power embedded systems since the accompanied high energy demand results in a short battery lifetime. Thus, we present a new approach for this problem setting using the same dataset and the same split resulting in a higher accuracy while requiring a much smaller network and less parameters.

2 Methodology

2.1 Inference

To improve the diagnosis within the GI tract, a tool for accurate self-localization of the capsule is presented. Since the energy limitation of such a small device needs to be considered, it is crucial to limit the size of the typical large deep neural network. The presented approach achieves this by improving the classification results of a relatively small CNN with subsequent time-series analysis.

CNNs have been successfully used for many different domains such as computer vision, speech and pattern recognition tasks [1, 8, 11] and thus, were employed for the classification task in this work. MobileNets [10] can be categorized as leight weight CNNs, which have been used for recognition tasks while being efficiently employed on mobile devices. Since a low model complexity is essential for the capsule application, the MobileNetV3-Small [9] was utilized and is in the following interchangeably referred to as CNN. For subsequent time-series analysis, a HMM was chosen, since the statistical model is well established in the context of time series data [20]. Due to the natural structure of the GI tract in humans, the order of states within a VCE is known. The capsule traverses the esophagus, stomach, small intestine and colon sequentially in every study. This inherent structure of visited locations can be directly encoded into the transition probabilities of the HMM. Finally, the predictions from the CNN are interpreted as the emissions of a HMM and the Viterbi algorithm [5] is used to compute the most likely sequence of exact locations, given the classifications of the CNN.

Fig. 1: Illustration of the presented approach (GI tract images from [2]).

Hence, the presented method for localizing the four different gastroenterology sections consists of two phases as depicted in Figure 1. For each patient, the CNN classifies chronologically received input data from the RI gastroenterology VCE dataset [2] into the four given classes. The respective output labels/predictions of the CNN are fed into a HMM, which uses the Viterbi algorithm for determining the most likely sequence of states given the observations from the CNN. With four hidden states and often much more than 10000 observations per patient, the size of the matrix storing the likelihood values for each hidden state and each observation has usually more than 40000 entries. However, the less storage is required, the more useful this method becomes for low power devices. Furthermore, as the decoding is performed backwards a larger matrix leads to an increasing delay in classification. Thus, to limit the size of the matrix, a sliding window of

size n was used to build the matrix with a predefined shape. After succeeding the n^{th} observation, for each new addition to the matrix the first column is removed, ensuring the predefined shape. The designated route the capsule moves along is known by the given anatomy of the GI tract. Therefore, specific assumptions can be made confidentially, e.g. the capsule cannot simply skip an organ, nor does the capsule typically move backwards to an already passed organ. To exploit this advantage of prior knowledge, the Viterbi decoding is used to detect the transitions for each organ by limiting the possible transition to the subsequent organ until a transition is detected.

2.2 HMM and Viterbi decoding

HMMs are popular tools in the context of time-dependent data, e.g. in pattern recognition tasks [21, 13, 25], characterized by low complexity compared to other models. The probabilistic modeling technique of a HMM assumes an underlying Markov chain, which describes the dynamic of the hidden states $\{S_1, \dots, S_n\}$. In the present setting, the exact location of the capsule is interpreted as the hidden state, which gives $\{S_1 = \text{Esophagus}, S_2 = \text{Stomach}, S_3 = \text{Small intestine}, S_4 = \text{Colon}\}$. At any time point t, an emission $X_t \in \{K_1, \dots, K_m\}$, corresponding to one of the locations as classified by the neural network, is observed and assumed to be sampled from the emission probabilities, which only depend on the hidden state S_t. The model is thus completely determined by the transition probabilities a_{ij} of the Markov chain S_t (as well as its initial distribution $\pi_i = P(X_1 = S_i)$ for $i \leq 4$) and the emission probabilities $b_j(k)$, which are given by the probabilities

$$a_{ij} = P(X_{t+1} = S_j | X_t = S_i), \quad b_j(k) = P(O_t = K_k | X_t = S_j),$$

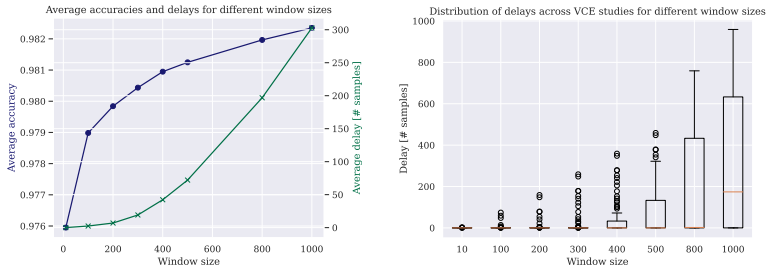
for $i, j \leq 4$ and $k \leq 4$, where O_t denotes the observation at time t [14, 4]. The objective of the model is then to infer the most likely hidden states, given observations $O = (O_1, \dots, O_t)$, which is effectively realized by the Viterbi algorithm [5]. The Viterbi algorithm then efficiently computes the most likely sequence of gastroenterologic states $X = (X_1, \dots, X_t)$, given the evaluations of the neural network $(O_1, \dots, O_t)$, namely

$$\arg \max_{X_1, \dots, X_t} P(X_1, \dots, X_t, O_1, ...O_t).$$

To determine good approximations for the transition and emission probabilities in this problem setting, a grid search was performed. For the diagonal and superdiagonal entries of the probability matrix, a defined number of values was tested as different combinations and for each variation the average accuracy computed for all patients (all other values were set to zero). The final probabilities were then chosen based on the obtained accuracies from the grid search and implemented for all following experiments. An additional metric is employed to evaluate the time lag between the first detection of an image originating from the small intestine and the actual passing of the capsule at this position. This delay arises due to the required backtrace of the matrix storing the log-likelihoods during the Viterbi decoding before the final classification can be performed.

3 Results and Discussion

To determine the window size used during Viterbi decoding for subsequent experiments, the average accuracy as well as the average delays of the Viterbi decoding after classification with the CNN+HMM combination are plotted over different window sizes (see Figure 2). A larger window size results in a higher accuracy and a larger delay, while a window size reduction leads to an accuracy loss but also a decrease of the delay. A reasonable tradeoff seems to be given by a window size of 300 samples, which was chosen for further experiments.



(a) Average delays and accuracies.

(b) Distribution of delays.

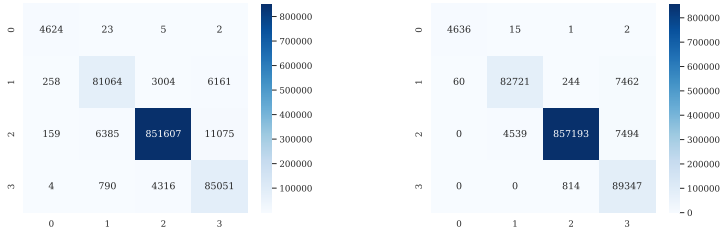
Fig. 2: Delays and accuracies for different window sizes sliding over the log-likelihood matrix of the Viterbi decoding.

Table 1 displays the results of the CNN+HMM combination in comparison to only using the CNN on different input image sizes (averaged describes the average over the values per patient). The $n \times n$ center of the input image was cropped and used as an input to explore the reduced complexity in terms of accuracy. This demonstrates that combining the CNN with time-series analysis can compensate a proportion of false classifications of the CNN and enhances its overall classification abilities notably. However, for all subsequent experiments an input image size of 320×320 was used for better comparability with the results of the original authors.

In all experiments, the MobileNetV3 was trained for 10 epochs on the RI training set with the AdamW optimizer and a learning rate of 0.001 with the HANNAH framework [6]. The presented approach for localization within the GI tract by combining classification with the CNN and the time-series analysis of the HMM achieved an accuracy of 98.04% with a window size of $w = 300$. This is an improvement compared to only applying the MobileNetV3 on its own (96.95%). The corresponding confusion matrices are shown in Figure 3, displaying the improved classification per class of the combination CNN+HMM (b) and compared to solely using the CNN (a). It becomes apparent that particularly

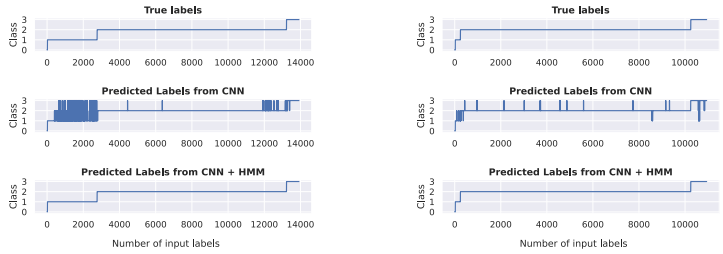
Table 1: Results of the CNN+HMM approach with different input image sizes.

Input size	64 × 64		120 × 120		320 × 320	
Metric	CNN	CNN+HMM	CNN	CNN+HMM	CNN	CNN+HMM
Accuracy [%]	90.60	96.16	93.94	97.37	96.95	98.04
Averaged MAE	0.1178	0.0560	0.07895	0.0406	0.0463	0.0350
Averaged R2-Score	0.1764	0.6889	0.4454	0.7819	0.7216	0.8077
Average Delay	–	19.11	–	17.87	–	19.19



(a) CNN output. (b) CNN + HMM output.

Fig. 3: Confusion matrices of the CNN output (a) and the CNN+HMM combination (b) (classes: esophagus (0), stomach (1), small intestine (2) and colon (3)).



(a) VCE study ID 94. (b) VCE study ID 95.

Fig. 4: Comparison of class predictions, exemplarily shown for two VCE studies.

the classification of stomach and colon images was notably enhanced (3004 vs. 244 images misclassified as small intestine and 4316 vs. 814 images misclassified as colon). Hence, even with low resolutions high accuracies were achieved (Table 1) and an overall improved self-localization was demonstrated (Figure 3). Importantly, this allows to use low resolutions within the sections outside of the small intestine while still providing a precise classification. Therefore, less en-

ergy is required within these sections and the energy can be either used for other tasks or simply leads to a longer battery lifetime. Table 2 extends the results by displaying additional metrics in comparison to the baseline [2]. This demonstrates the size reduction of the neural network to $\approx$ 1M parameters compared to the baseline model [2] with $\approx$ 56M parameters (both computed with 32-bit floating-point values).

Exemplarily, in Figure 4, for two patients, the true labels (top row, corresponding to the perfect solution) of the captured images over time are shown in comparison to the predicted labels from the CNN only (middle row) and finally the predicted labels from the HMM which further processed the output from the CNN (last row). While the CNN still presents some misclassifications, the HMM is able to capture and correct false predictions from the CNN to some extend, resulting in a more similar depiction compared to the true labels.

Table 2: Results for the presented approach in comparison to the baseline results (Mean values over all 85 tested VCE studies).

Metric	MobilenetV3 + HMM ($w = 300$)	MobilenetV3	Baseline [2]
Accuracy [%]	98.04	96.95	97.1
Number of Parameters	$\approx$ 1M	$\approx$ 1M	$\approx$ 56M
Averaged MAE	0.0350	0.0463	–
Averaged R2-Score	0.8077	0.7216	–
Average Delay (# Frames)	19.19	–	–

The accuracies of class prediction achieved with the CNN compared to the combinatorial approach over all patient VCE studies are visualized in Figure 5. The CNN+HMM combination achieved superior results compared to the CNN alone for almost all patient studies. Exemplarily, two of the outliers are observed more closely to understand the incidences of worse performance with the combinatorial approach (Figure 6). Presented in Figure 6a, one VCE study shows poor results for both approaches. The CNN misclassifies the majority of the images achieving an accuracy of only 25.20%. Subsequently, as the preceding classification of the CNN is mostly incorrect, the Viterbi decoding cannot classify the error-prone labels correctly received from the CNN resulting in a very early misclassification as colon. Since the HMM cannot take a step back and the CNN provides such an error-prone output, the remaining images are also classified as colon resulting in an accuracy of 5.93%. For one VCE study (Figure 6b) overall good results can be achieved, but classification with the combination of CNN+HMM showed a slightly worse result than classifying with the CNN only. It becomes apparent that the CNN has trouble classifying images near the transition of small intestine and colon leading to a delayed transition detection by the HMM. It is expected that a large proportion of misclassifications can be avoided with time-dependent transition probabilities.

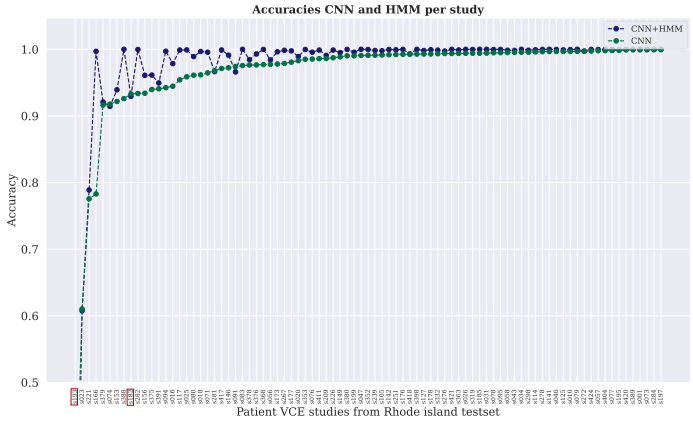
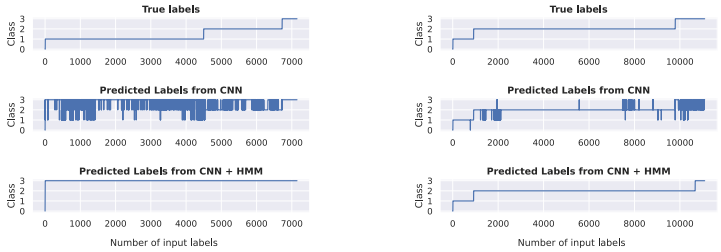


Fig. 5: Accuracies of the CNN compared to the combinatorial approach CNN+HMM. Marked in red are two studies with worse results if the combination is used, more details can be found in Figure 6a and Figure 6b.



(a) VCE study 193.

(b) VCE study 182.

Fig. 6: Comparison of class predictions for two outlier VCE studies.

4 Conclusion

A pipeline to accurately classify the current gastroenterologic section of VCE images was proposed. The combination of a CNN followed by time-series analysis can automatically classify the present section of the capsule achieving an accuracy of 98.04%. Limited by the small size of the embedded device in practical use, considering the given energy constraints is crucial. The presented approach requires only $\approx$ 1M parameters while providing a higher accuracy than the current baseline. This approach results in an energy reduction which potentially provides more options on when and how relevant images are captured.

References

1. Abdel-Hamid, O., Mohamed, A.r., Jiang, H., Deng, L., Penn, G., Yu, D.: Convolutional neural networks for speech recognition. IEEE/ACM Transactions on audio, speech, and language processing **22**(10), 1533–1545 (2014)
2. Charoen, A., Guo, A., Fangsaard, P., Taweechainaruemitr, S., Wiwatwattana, N., Charoenpong, T., Rich, H.G.: Rhode island gastroenterology video capsule endoscopy data set. Scientific Data **9**(1), 602 (2022)
3. Costamagna, G., Shah, S.K., Riccioni, M.E., Foschia, F., Mutignani, M., Perri, V., Vecchioli, A., Brizi, M.G., Picciocchi, A., Marano, P.: A prospective trial comparing small bowel radiographs and video capsule endoscopy for suspected small bowel disease. Gastroenterology **123**(4), 999–1005 (2002)
4. Eddy, S.R.: Hidden markov models. Current opinion in structural biology **6**(3), 361–365 (1996)
5. Forney, G.D.: The viterbi algorithm. Proceedings of the IEEE **61**(3), 268–278 (1973)
6. Gerum, C., Frischknecht, A., Hald, T., Bernardo, P.P., Lübeck, K., Bringmann, O.: Hardware accelerator and neural network co-optimization for ultra-low-power audio processing devices. arXiv preprint arXiv:2209.03807 (2022)
7. Goh, S.T., Zekavat, S.A., Pahlavan, K.: Doa-based endoscopy capsule localization and orientation estimation via unscented kalman filter. IEEE Sensors Journal **14**(11), 3819–3829 (2014)
8. 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)
9. Howard, A., Sandler, M., Chu, G., Chen, L.C., Chen, B., Tan, M., Wang, W., Zhu, Y., Pang, R., Vasudevan, V., et al.: Searching for mobilenetv3. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 1314–1324 (2019)
10. Howard, A.G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., Adam, H.: Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861 (2017)
11. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4700–4708 (2017)
12. Iddan, G., Meron, G., Glukhovsky, A., Swain, P.: Wireless capsule endoscopy. Nature **405**(6785), 417–417 (2000)
13. Kenny, P., Lennig, M., Mermelstein, P.: A linear predictive hmm for vector-valued observations with applications to speech recognition. IEEE Transactions on Acoustics, Speech, and Signal Processing **38**(2), 220–225 (1990)
14. Manning, C., Schutze, H.: Foundations of statistical natural language processing. MIT press (1999)
15. Marya, N., Karellas, A., Foley, A., Roychowdhury, A., Cave, D.: Computerized 3-dimensional localization of a video capsule in the abdominal cavity: validation by digital radiography. Gastrointestinal endoscopy **79**(4), 669–674 (2014)
16. Mateen, H., Basar, R., Ahmed, A.U., Ahmad, M.Y.: Localization of wireless capsule endoscope: A systematic review. IEEE Sensors Journal **17**(5), 1197–1206 (2017)
17. McLaughlin, P.D., Maher, M.M.: Primary malignant diseases of the small intestine. American Journal of Roentgenology **201**(1), W9–W14 (2013)
18. Monteiro, S., de Castro, F.D., Carvalho, P.B., Moreira, M.J., Rosa, B., Cotter, J.: Pillcam® sb3 capsule: Does the increased frame rate eliminate the risk of missing lesions? World Journal of Gastroenterology **22**(10), 3066 (2016)

19. Pham, D.M., Aziz, S.M.: A real-time localization system for an endoscopic capsule. In: 2014 IEEE Ninth International Conference on Intelligent Sensors, Sensor Networks and Information Processing (ISSNIP). pp. 1–6. IEEE (2014)
20. Rabiner, L., Juang, B.: An introduction to hidden markov models. ieee assp magazine **3**(1), 4–16 (1986)
21. Rabiner, L.R.: A tutorial on hidden markov models and selected applications in speech recognition. Proceedings of the IEEE **77**(2), 257–286 (1989)
22. Smedsrud, P.H., Thambawita, V., Hicks, S.A., Gjestang, H., Nedrejord, O.O., Næss, E., Borgli, H., Jha, D., Berstad, T.J.D., Eskeland, S.L., et al.: Kvasir-capsule, a video capsule endoscopy dataset. Scientific Data **8**(1), 142 (2021)
23. Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A.: Inception-v4, inception-resnet and the impact of residual connections on learning. In: Proceedings of the AAAI conference on artificial intelligence. vol. 31 (2017)
24. Thomson, A., Keelan, M., Thiesen, A., Clandinin, M., Ropeleski, M., Wild, G.: Small bowel review: diseases of the small intestine. Digestive diseases and sciences **46**, 2555–2566 (2001)
25. Trentin, E., Gori, M.: Robust combination of neural networks and hidden markov models for speech recognition. IEEE Transactions on Neural Networks **14**(6), 1519–1531 (2003)
26. Yim, S., Sitti, M.: 3-d localization method for a magnetically actuated soft capsule endoscope and its applications. IEEE Transactions on Robotics **29**(5), 1139–1151 (2013)
27. Zhang, L., Zhu, Y., Mo, T., Hou, J., Rong, G.: Design and implementation of 3d positioning algorithms based on rf signal radiation patterns for in vivo micro-robot. In: 2010 International Conference on Body Sensor Networks. pp. 255–260. IEEE (2010)
28. Zhou, M., Bao, G., Pahlavan, K.: Measurement of motion detection of wireless capsule endoscope inside large intestine. In: 2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society. pp. 5591–5594. IEEE (2014)

Enhanced Anomaly Detection for Capsule Endoscopy Using Ensemble Learning Strategies*

Julia Werner[1], Christoph Gerum[1], Jörg Nick[2], Maxime Le Floch[3,4], Franz Brinkmann[3,4], Jochen Hampe[3,4], and Oliver Bringmann[1]

***Abstract*— Capsule endoscopy is a method to capture images of the gastrointestinal tract and screen for diseases which might remain hidden if investigated with standard endoscopes. Due to the limited size of a video capsule, embedding AI models directly into the capsule demands careful consideration of the model size and thus complicates anomaly detection in this field. Furthermore, the scarcity of available data in this domain poses an ongoing challenge to achieving effective anomaly detection.**

Thus, this work introduces an ensemble strategy to address this challenge in anomaly detection tasks in video capsule endoscopies, requiring only a small number of individual neural networks during both the training and inference phases. Ensemble learning combines the predictions of multiple independently trained neural networks. This has shown to be highly effective in enhancing both the accuracy and robustness of machine learning models. However, this comes at the cost of higher memory usage and increased computational effort, which quickly becomes prohibitive in many real-world applications. Instead of applying the same training algorithm to each individual network, we propose using various loss functions, drawn from the anomaly detection field, to train each network. The methods are validated on the two largest publicly available datasets for video capsule endoscopy images, the Galar and the Kvasir-Capsule dataset. We achieve an AUC score of 76.86% on the Kvasir-Capsule and an AUC score of 76.98% on the Galar dataset. Our approach outperforms current baselines with significantly fewer parameters across all models, which is a crucial step towards incorporating artificial intelligence into capsule endoscopies.

I. Introduction

Gastrointestinal (GI) diseases can affect the life quality severely. While some are easier to detect, others remain hidden until the consequences are too severe. Regions such as the mouth, esophagus, stomach as well as the colon can be successfully evaluated by standard techniques, such as a gastroscopy or colonoscopy [27], [33]. However, the small intestine is predominantly inaccessible by such endoscopes. A minimal invasive method that has been emerged since the early 2000s is the capsule endoscopy, which can visualize this part of the GI tract [9], [11], [18]. With a video capsule endoscopy (VCE), a patient ingests a small pill-sized capsule comprising of an integrated camera and a light-emitting diode. As it traverses the GI tract through peristalsis, the capsule transmits recorded images to an electronic device [22]. Afterwards, this enables the identification of anomalies within the duodenum, jejunum and the ileum, which correspond to the distal first section, the midsection and the last section of the small intestine, respectively [19]. However, given the limited size of such video capsule, incorporating artificial intelligence (AI) directly into the capsule requires careful consideration of the overall model size and thus complicates anomaly detection on-site. Additionally, due to the limited data availability in this field, successful anomaly detection remains challenging in general. A classical approach to improve the performance of machine learning methods is ensemble learning [10], [25], [31]. It integrates multiple models and hereby balances the weaknesses of individual models.

To leverage this, in this work, ensemble learning is applied to VCE data by combining different anomaly detection methods while considering the overall model size compared to the current state-of-the-art methods in this field. This provides a starting point to construct hardware-aware models enabling on-site anomaly detection. Inference on the hardware of the capsule endoscopy is particularly desirable for adaptive sampling rates and transmission. Upon anomaly detection, only the picture of interest should be transmitted for further examination by specialized physicians. Through this on-site evaluation, the required energy of such a capsule can be drastically reduced since the costly transmission of unproblematic images is omitted. Furthermore, this would allow the determination of anomalies in real-time and for example, to act immediately upon detection by increasing the resolution or the frame rate on-site at the region of interest. As a first step, this work transfers well-established anomaly detection techniques to this specific medical setting and explores reliable practices to combine those. Within this process, the total number of parameters of the proposed models as well as the feasibility to implement this in hardware, is kept in mind.

Our contribution. In this paper, we describe a methodology for the construction of hardware-efficient ensemble learning models for anomaly detection tasks. We achieve this goal by formulating different anomaly detection algorithms based on the hardware-efficient MobileNet architecture, which are then combined in a simple ensemble method procedure to construct efficient anomaly detection algorithms. The proposed methodology produces models that outperform state-of-the-art models on the two largest video capsule endoscopy datasets, while requiring only a fraction

*This work has been partly funded by the German Federal Ministry of Education and Research (BMBF) in the project MEDGE (16ME0530).

[1] Department of Computer Science, University of Tübingen, Tübingen, Germany

[2] Seminar of Applied Mathematics, ETH Zürich, Zürich, Switzerland

[3] Else Kröner Fresenius Center for Digital Health, TU Dresden, Dresden, Germany

[4] Department of Medicine I, University Hospital Dresden, TU Dresden, Dresden, Germany

of the storage memory cost compared to the state-of-the-art methods in this medical field. With these results, we aim to provide a key contribution towards the on-edge application of machine learning models for capsule endoscopies.

II. Background and Related Work

In this section, the main components of the present work are described, namely background on ensemble learning, anomaly detection methods as well as the current state of publicly available datasets comprising of video capsule endoscopy studies.

A. Ensemble Learning

Ensemble learning can be employed to combine various machine learning methods and thereby, to better capture the underlying structure of data [10], [25], [31]. It has been used for anomaly detection in a number of fields, for example in network security [28], [31] or in medicine for neurocognitive disorder detection based on MRI datasets [26].

Models based on ensemble learning often provide superior performance compared to their individual building blocks - at the cost of additional parameters and significant more operations during inference. Following [15], different techniques have been presented to accelerate the inference of ensemble models, see e.g. [32]. Here, we aim to alter the training methods on the same architecture and construct methods that require only very few neural network evaluations during inference. We therefore do not require additional model order reduction techniques at this stage, although these ideas might provide avenues for future research.

B. Anomaly Detection

Anomaly detection remains a critical challenge in a wide range of applications, such as cyber security, network intrusion detection or health care [6]. Many methodologies evolved across various fields of machine learning that have been proven effective for outlier detection. In the field of supervised machine learning, models are trained on labeled data only. However, this relies on the availability of preferably large labeled datasets and especially in the field of medicine, this is less suitable due to the scarcity of labeled data. Unsupervised techniques form an alternative by exploiting the existence of unlabeled data, e.g. by employing deep autoencoders [37]. Semi-supervised techniques additionally make use of unlabeled as well as normal data [1]. In this work, we aim to harness the strengths of each area (supervised, unsupervised, semi-supervised) to maximize the use of the limited available data, while keeping in mind that only small networks can ultimately be deployed on hardware for this application.

C. Hardware-Aware Machine Learning Methods

The implementation of machine learning methods on edge devices remains a challenge, in particular when strong energy constraints are imposed on the available hardware. The development of hardware aware machine learning methods is therefore an active field of research and continuously evolving. A successfull family of network architectures based on depthwise separable convolutions is the MobileNet architecture [17], which was explicitly designed for embedded and mobile devices. In many real-world applications, these networks have demonstrated competitiveness with much larger models while using significantly fewer parameters. Finally, low-power hardware accelerators, explicitly designed for the efficient inference of neural network architectures, have been shown to significantly reduce power usage and inference time for restricted neural network architectures [4].

In this paper, we combine the progress made in the development of these architectures with anomaly detection and ensemble learning techniques, to construct hardware aware methods for video capsule endoscopies. In particular, for the medical investigation of the gastrointestinal tract only methods suitable for real-time low-power hardware architectures can be considered when neural networks are transferred to the real-world application.

D. Datasets for Capsule Endoscopy

The presented methods have been developed for the following two datasets, which are the largest publicly available video capsule datasets. In general, video capsule endoscopy datasets are sparse but remain essential if one wants to target pathological frames within capsule endoscopies.

Kvasir-Capsule. In 2021, the video capsule endoscopy dataset Kvasir-Capsule Dataset [29], has been published. To our knowledge, representing the only publicly available large video capsule dataset for several years with a total of $47,238$ labeled and $4,694,266$ unlabeled frames. As mentioned by the authors, it is crucial to use the official split as published [29]. Simply splitting all available images does not ensure that images from the same patient appear only in the training but not in the test set. Due to the close similarity of frames, one would need to perform splits on the patient studies. Thus, we only compare results from this work with research that stuck to the official given splits or specifically state that they performed a split patient-wise.

Galar Dataset. Recently, in 2024, a new large multi-label video capsule endoscopy dataset was published: Galar [20], which provides a total of $3,513,539$ labeled images from 80 patient VCE studies in total. This dataset contains annotated images with a variety of classes, including anatomical as well as pathological findings.

In the following, we propose a combination of methods leading to improved results for both datasets while using notably smaller models. Employing smaller networks is crucial if one aims for execution on low-power hardware architectures as employed for capsule endoscopy, which is predominantly limited by its capsule size. In this work, we validate and compare our methods on those two largest publicly available VCE datasets.

III. Methodology and Experiments

Medical datasets are expensive to label and therefore often contain large amounts of unlabeled data. Nevertheless,

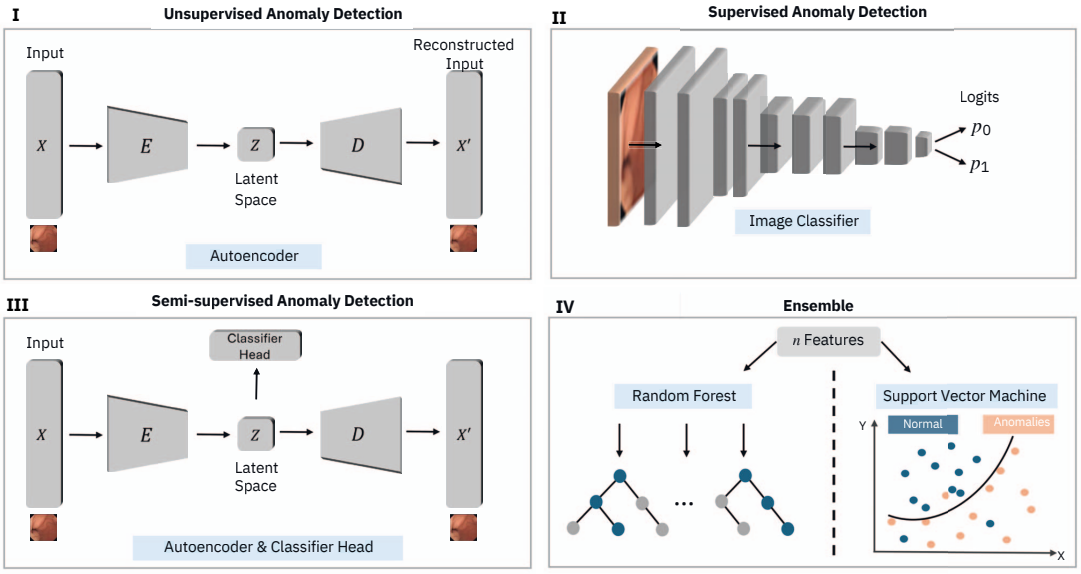


Fig. 1. Overview of the ensemble learning strategy for Anomaly detection, consisting of an unsupervised (I), a supervised (II) and a semi-supervised (III) classification approach, followed by an ensemble model constituting of either a random forest model or a SVM (IV). I, II and III are all based on the same network architecture.

significant progress has been made in the acquisition of labeled medical data, as is demonstrated by the two capsule endoscopy datasets treated in this paper.

Our method therefore makes use of the substantial amount of labeled data present as well as the vast amount of unlabeled data available. Consequently, we combine three different types of classifiers: Supervised anomaly detection, Semi-supervised anomaly detection and unsupervised anomaly detection.

Figure 1 illustrates the overall classification approach. It comprises four parts, (I) an autoencoder, representing the unsupervised technique, (II) a standard image classifier drawn from the supervised methods and (III) an autoencoder in combination with an additional classification head, which forms the semi-supervised approach. The final ensemble model (IV), consisting of either a Random Forest model or a Support Vector Machine (SVM), combines the prediction of each method and returns the final evaluation.

A. Datasets and Preprocessing

The main goal of this proposed pipeline is the anomaly detection of VCE images, especially in the region of interest, the small intestine. To achieve this, the classes of the Kvasir-Capsule dataset were categorized into two classes, normal and anomaly, for binary classification:

- **Normal:** Pylorus, Reduced Mucosal View, Ileo-cecal valve, Normal Clean Mucosa
- **Anomaly**: Angiectasia, Blood-fresh, Foreign Bodies, Ulcer, Erosion, Lymphangiectasia.

Figure 2 visualizes the huge class imbalances within the Kvasir-Capsule dataset. While there are approximately $4,7$ M unlabeled frames, within the $\approx 47,000$ labeled frames, only about 9% are pathological images. This highlights the importance of employing machine learning methods that specifically target such outliers.

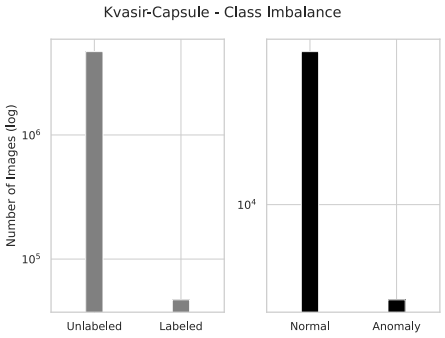


Fig. 2. Visualization of the number of images available in the Kvasir-Capsule dataset (unlabeled vs. labeled and normal vs. anomaly).

The Galar dataset as originally published is partitioned into different splits for each individual pathology but not for multiple pathological classes jointly. In order to perform anomaly detection, new splits for the training and validation sets were generated patient-wise, assembling the corresponding following pathology classes drawn from the small intestine only:

- **Normal:** Normal Clean Mucosa
- **Anomaly**: Polyp, Blood, Active Bleeding, Angiectasia,

Erosion, Erythema, Ulcer.

The test set still comprises the same patient studies as originally published (patient IDs $61-80$). The remaining patient studies (ID $1-60$) were split in a $80:20$ (train : val) ratio.

Depending on the camera, the videos were originally recorded with a varying resolution ranging from 256×256 to 512×512 [29] for Kvasir-Capsule and from 336×336 to 576×576 pixels for the Galar dataset [20], which we initially adapted. However, since reducing the image sizes can lead to a model with lower complexity, we reduced the resolution for training the images to 224×224 as also employed by [20], while keeping the original image sizes for testing. We found that reducing the image size to 224×224 does not impair the overall model accuracy. For data augmentation, random rotation, random vertical and horizontal flip as well as random erasing of small parts of images was applied.

B. Autoencoder - Unsupervised and Semi-Supervised Anomaly Detection

As demonstrated in the past, autoencoders do not only have remarkable reconstruction capabilities but are furthermore effective for outlier detection [2], [12], [36]. They are also beneficial in a semi-supervised setting, by enabling the incorporation of unlabeled data. Since the given Kvasir-Capsule dataset has millions of unlabeled data, we employ autoencoders to leverage this.

An autoencoder A is trained such that the decoder D learns to reconstruct a given input X which is encoded by an encoder E. Thus, the following optimization problem needs to be solved

$$\min_{D,E} = \|X - D(E(X))\|, \tag{1}$$

with $\|\cdot\|$ as the $l2$-norm [36].

The mean squared error (MSE) [14] was employed as a reconstruction loss in all autoencoder experiments:

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2, \tag{2}$$

and is directly used as a feature in the ensemble model. One special characteristic in this setting is that the autoencoder is based on the same network architecture as has been used for the image classifier. More precisely, the encoder consists of the MobilNetV3-Small architecture. According to this, the decoder was generated with the PyTorch framework and the pretrained model fine-tuned for additional 15 epochs. Data augmentation is employed to reduce overfitting. Hence, by training the autoencoder only with unlabeled or normal samples, we tried to emphasize a good reconstruction capability of normal samples and an inferior reconstruction performance for unseen anomalies. In the semi-supervised setting, for training, unlabeled as well as labeled images were used for the Kvasir-Capsule dataset. The Galar dataset only comprises labeled frames, thus, only labeled data was incorporated. Additionally, the compressed data from the latent space was processed with a small classifier head and its final prediction also used as a feature for the ensemble model.

C. Image Classification - Supervised Anomaly Detection

MobileNets have been successfully applied to image classification and recognition tasks, with the added advantage of being highly efficient on mobile as well as embedded devices [8], [16], [17]. This makes them an ideal solution for the given problem, as they are well-suited for low-power hardware while still delivering strong classification performance. In the past, others have generated FPGA-based accelerators specifically designed for this architecture and proven its efficiency [21], [34]. Therefore, as the most light-weight approach, standard image classification was performed by fine-tuning a pretrained MobileNetV3 for 15 epochs with the Adam optimizer, using PyTorch [3]. Additionally, a weighted sampler was employed to address the omnipresent class imbalances of these two datasets. In this supervised setting, training was only performed with labeled anomaly and normal VCE images by employing the cross-entropy (CE) [3], [35] loss

$$\text{CE} = -\frac{1}{N}\sum_{i=1}^{N} y_i \log(p(y_i)) + (1 - y_i)\log(1 - p(y_i)). \tag{3}$$

D. Ensemble Model

The key concept of the presented methodology is to provide an ensemble method that yields an overall improved prediction performance and therefore outperforms the individual methods by combining all given models. As described by [25], for a dataset with m features and n samples $D = \{(x_i, y_i) \mid x_i \in \mathbb{R}^m, y_i, \in \{0, 1\}, i = 1, \ldots, n\}$, the classification of an ensemble model ϕ with K base learners $\{f_0, f_1, \ldots, f_k\}$ can be described by the following function:

$$\hat{y}_i = \phi(x_i) = H(f_0, f_1, ..., f_k),$$

with $\hat{y}_i \in \mathbb{Z}$ for classification problems. For our setting, we apply $k = 3$ different base learners (autoencoder, autoencoder with classification head and image classifier) and two different ensemble models ϕ (Random Forest Classifier [5] or a SVM [7], [13], with $\hat{y}_i \in \{0, 1\}$. A random subset of the training and validation sets were used to perform a random search and find the best parameters for each ensemble model and to finally train those models. For final evaluation, the test set was used.

IV. Results

The results of the experiments performed on the Kvasir-Capsule dataset are listed in Table I. For the 12-class problems treated in recent research, the macro-average, the micro-average and weighted metrics are shown, if accessible. If someone aims to find all pathological images, the macro-average weighted metrics are considered to be the most relevant ones for the multi-class settings with large class imbalances as the different classes are weighted equally. For the micro-average and the weighted metrics, the larger

class is given more weight, which is not of interest in anomaly detection since the larger class typically consists of the normal samples. However, for the purpose of completeness, both metrics are shown, if accessible. For binary-class problems, such distinction is not relevant, which impedes the comparison across these studies. Thus, for the final comparison, we consider only the results presented below the dashed line. Nevertheless, those baseline results are also shown in order to provide an overview of essential results in this field.

TABLE I
RESULTS OF THE ENSEMBLE MODEL ON THE KVASIR-CAPSULE DATASET COMPARED TO THE BASELINE RESULTS.

Method	AUC	Recall	Accuracy	F1 Score	MCC	Precision
Baseline Results						
DenseNet161, Baseline [29] - macro	–	28.12	–	25.60	43.29	29.94
DenseNet161, Baseline [29] - micro	–	73.66	–	73.66	43.29	73.66
FocalConvNet [30] - macro	–	27.45	63.73	21.78	29.64	24.38
FocalConvNet [30] - weighted	–	63.73	63.73	67.34	29.64	75.57
ViT [23] - weighted	57.0	71.56	71.56	71.56	37.05	68.41
ResNet152 and OneClassSVM [24]	-	56.00	-	50.00	-	55.00
ResNet152 and XGBoost [24]	-	56.00	-	57.00	-	73.00
This Work						
Image Classifier (CLF)	73.42	51.96	91.65	48.30	43.91	45.12
Autoencoder (AE)	64.70	40.90	84.94	28.96	22.54	22.42
Ensemble SVM, AE, CLF	76.80	61.59	89.72	47.36	43.43	38.47
Ensemble RF, AE, CLF	76.86	60.65	90.64	49.32	45.33	41.56

The authors from the Kvasir-Capsule dataset [29] yield a macro-average precision of 29.94%, a recall of 28.12% and a F1-score of 25.60% with a DensNet-161. Srivastava et al. [30] achieved an accuracy of 63.73% and a macro-average recall of 27.45% with a FocalConvNet including depth-wise separable convolutions followed by a GeLU activation layer. From those authors listed above, only de Sá et al [24] also performed a binary study, splitting the Kvasir-Capsule dataset into binary classes in a similar way as conducted in this work and is therefore ideally suited for comparison. They trained a ResNet152 and fine-tuned it with XGBoost and an OneClassSVM. They report a recall of 56%, a F1-score of 57% and a precision of 73% while requiring about ≈ 60 million parameters.

Table I demonstrates that both ensemble models outperform the individual methods in terms of the AUC score (76.86%), sensitivity (60.65%) and Matthew's correlation coefficient (MCC) (45.33%). The accuracy as well as the precision is higher if only the image classifier is used. However, considering the low sensitivity, this is most likely due to a large fraction of true negatives while not detecting many true positives. Overall, both ensemble models involving either the SVM or Random Forest perform well compared to [24] by detecting a larger fraction of true positives ($\approx 61\%$ vs. 56%). This partially probably comes at the cost of a lower precision compared to [24]. Importantly, while demonstrating a good performance, the total model sizes of all presented models is drastically reduced from around 60 million to a maximum of 4 million parameters.

Following this, the proposed approach was validated on anomalies in the small intestine from the Galar dataset. Since this dataset has just been recently published, there are no results published targeting the pathologies jointly. The image classifier yields the highest scores in terms of accuracy, F1 score, MCC score and precision. From our proposed models, the ensemble model including the SVM attains the highest AUC score (76.98%). The ensemble model including the random forest yields the highest recall (86.06%), but has an inferior accuracy compared to the image classifier. Interestingly, in contrast to the Kvasir-Capsule dataset, the Random Forest ensemble model is inferior compared to the SVM ensemble model on the Galar dataset. Overall, the image classifier shows a strong performance while the autoencoder exhibits inferior results. One possible reason, in comparison to the Kvasir-Capsule dataset, is that the availability of a substantial amount of labeled data allows the image classifier to be well-trained independently. However, without any unlabeled data, the strengths of the autoencoder can not be fully exploited. Building on the successful classification of individual anomaly tasks in [20], we present the first results on anomaly detection with the majority anomaly classes pooled together from this dataset, offering a potential foundation for future work.

TABLE II
RESULTS OF THE ENSEMBLE MODEL ON THE GALAR DATASET.

Method	AUC	Recall	Accuracy	F1 Score	MCC	Precision
Image Classifier (CLF)	74.94	60.88	87.28	37.01	34.45	26.59
Autoencoder (AE)	45.83	28.94	60.65	8.28	-4.15	4.83
Ensemble SVM, AE, CLF	76.98	80.58	73.83	27.44	28.29	16.53
Ensemble RF, AE, CLF	68.34	86.06	52.80	18.3	17.62	10.24

Additionally, we explored how well the individual pathologies were detected by plotting the proportion of correctly labeled classes for both datasets for one of the ensemble models exemplarily. For the Kvasir-Capsule dataset (Figure 3), among the normal samples, the pylorus class has the highest misclassification rate. Among the anomalies, the classes lymphangiectasia and erosion exhibited the highest error rates. The misclassification of the pylorus might be explained with the occasional occurence of anomalies such as erosions or redness around the pylorus region which have similarities with the small intestine pathologies and thus, might be recognized as such by the classifier. Figure 4 reveals huge differences in the proportion of correctly labeled classes for the galar dataset. While blood, active bleeding and angiectasias are surprisingly well detected, polyps and erosions are mostly not correctly classified. It's possible that these well-detected classes share certain properties that the ensemble identifies as anomalous.

The partial contribution of the individual components of the ensemble model and its overall advantage were particularly evident for the Kvasir-Capsule dataset. Thus, for this dataset, the classification results of the ensemble model including the random forest are more closely evaluated by plotting the distance of the logits of the image classifier over the

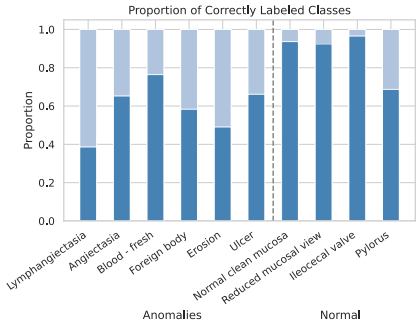


Fig. 3. Proportion of correctly labeled classes of the Kvasir-Capsule dataset if evaluated with the ensemble model including the Random Forest classifier, the autoencoder and the image classifier.

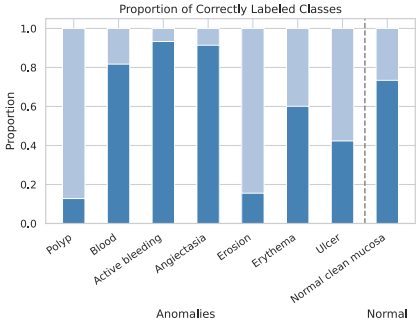


Fig. 4. Proportion of correctly labeled classes of the Galar dataset if evaluated with the ensemble model including a SVM, the autoencoder and the image classifier .

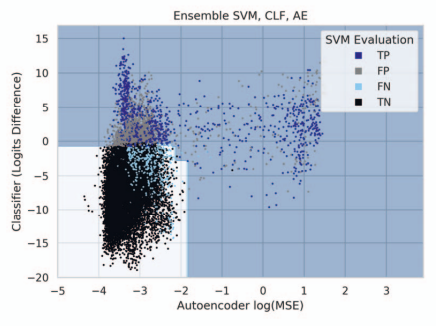


Fig. 5. Classification results of the ensemble model including the Random Forest classifier. The logits distance of the image classifier is plotted over the $\log(\text{MSE})$ of the autoencoder and the true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN) labeled by color as evaluated by the ensemble model. The blue area indicates the region in which the ensemble model classifies data samples as anomalous.

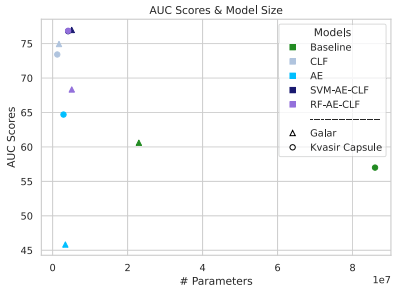


Fig. 6. AUC scores shown in comparison to the number of parameters needed for each model evaluated using both datasets: Galar and Kvasir-Capsule (Please note that the SVM-AE-CLF marking for Kvasir-Capsule is overlayed by the RF-AE-CLF marking point).

$\log(\text{MSE})$ outputted by the autoencoder and finally marked as evaluated by the Random Forest model in Figure 5. This shows, besides a large fraction of correctly classified true negatives, also a visible amount of false positives. However, detecting a large amount of true positives on the cost of some false positives is accepted in this setting because finding a majority of anomalies is prioritized. Furthermore, the shift of anomalies to the right side of the plot reveals that the autoencoder exhibits a higher loss on anomalies than on normal samples, as anticipated. We can further observe that the ensemble model recognizes that a larger MSE correlates with a higher possibility of actually being an anomaly.

Figure 6 visualizes the classification performance of each model and for each dataset while comparing the number of parameters needed for each model. The ensemble model including a SVM demonstrates superior performance on the Galar dataset. For Kvasir-Capsule, the Random Forest ensemble model yields the best results. Importantly, all implemented methods need less parameters than the baselines while resulting in a higher AUC score. More precisely, on the Kvasir-Capsule dataset we ultimately used around 4 million parameters instead of the 86 million parameters of the ViT from [23]. For the galar dataset, [20] performed all experiments with a ResNet50 with around 25 million [20] for single task pathology problems. We were able to reduce the number of parameters to 5 million and additionally perform anomaly detection on multiple pathologies.

V. Conclusion

In this paper, we developed hardware-aware ensemble learning methods for anomaly detection and validated them on a critical real-world application: video capsule endoscopy (VCE). Previous studies have shown that anomaly detection with the two largest VCE datasets presents significant challenges. By using the same network architecture as the backbone for each component of the ensemble and constraining the total number of parameters, we produced significantly smaller models with enhanced classification performance compared to state-of-the-art models. This is a crucial step towards a suitable AI model capable of anomaly detection for capsule endoscopies. Looking ahead, larger labeled datasets of pathological images from the small intestine are essential, as the current scarcity of data continues to impede progress in anomaly detection and remains a major bottleneck.

REFERENCES

[1] Akcay, S., Atapour-Abarghouei, A., Breckon, T.P.: Ganomaly: Semi-supervised anomaly detection via adversarial training. In: Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part III 14. pp. 622–637. Springer (2019)

[2] An, J., Cho, S.: Variational autoencoder based anomaly detection using reconstruction probability. Special lecture on IE **2**(1), 1–18 (2015)

[3] Ansel, J., Yang, E., He, H., Gimelshein, N., Jain, A., Voznesensky, M., Bao, B., Bell, P., Berard, D., Burovski, E., Chauhan, G., Chourdia, A., Constable, W., Desmaison, A., DeVito, Z., Ellison, E., Feng, W., Gong, J., Gschwind, M., Hirsh, B., Huang, S., Kalambarkar, K., Kirsch, L., Lazos, M., Lezcano, M., Liang, Y., Liang, J., Lu, Y., Luk, C., Maher, B., Pan, Y., Puhrsch, C., Reso, M., Saroufim, M., Siraichi, M.Y., Suk, H., Suo, M., Tillet, P., Wang, E., Wang, X., Wen, W., Zhang, S., Zhao, X., Zhou, K., Zou, R., Mathews, A., Chanan, G., Wu, P., Chintala, S.: PyTorch 2: Faster Machine Learning Through Dynamic Python Bytecode Transformation and Graph Compilation. In: 29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2 (ASPLOS '24). ACM (Apr 2024). https://doi.org/10.1145/3620665.3640366, https://pytorch.org/assets/pytorch2-2.pdf

[4] Bernardo, P.P., Gerum, C., Frischknecht, A., Lübeck, K., Bringmann, O.: Ultratrail: A configurable ultralow-power tc-resnet ai accelerator for efficient keyword spotting. IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems **39**(11), 4240–4251 (2020)

[5] Breiman, L.: Random forests mach learn 45 (1): 5–32 (2001)

[6] Chandola, V., Banerjee, A., Kumar, V.: Anomaly detection: A survey. ACM computing surveys (CSUR) **41**(3), 1–58 (2009)

[7] Chen, P.H., Lin, C.J., Schölkopf, B.: A tutorial on ν-support vector machines. Applied Stochastic Models in Business and Industry **21**(2), 111–136 (2005)

[8] Chiu, Y.C., Tsai, C.Y., Ruan, M.D., Shen, G.Y., Lee, T.T.: Mobilenet-ssdv2: An improved object detection model for embedded systems. In: 2020 International conference on system science and engineering (ICSSE). pp. 1–5. IEEE (2020)

[9] Costamagna, G., Shah, S.K., Riccioni, M.E., Foschia, F., Mutignani, M., Perri, V., Vecchioli, A., Brizi, M.G., Picciocchi, A., Marano, P.: A prospective trial comparing small bowel radiographs and video capsule endoscopy for suspected small bowel disease. Gastroenterology **123**(4), 999–1005 (2002)

[10] Dong, X., Yu, Z., Cao, W., Shi, Y., Ma, Q.: A survey on ensemble learning. Frontiers of Computer Science **14**, 241–258 (2020)

[11] Enns, R.A., Hookey, L., Armstrong, D., Bernstein, C.N., Heitman, S.J., Teshima, C., Leontiadis, G.I., Tse, F., Sadowski, D.: Clinical practice guidelines for the use of video capsule endoscopy. Gastroenterology **152**(3), 497–514 (2017)

[12] Gong, D., Liu, L., Le, V., Saha, B., Mansour, M.R., Venkatesh, S., Hengel, A.v.d.: Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 1705–1714 (2019)

[13] Hearst, M.A., Dumais, S.T., Osuna, E., Platt, J., Scholkopf, B.: Support vector machines. IEEE Intelligent Systems and their applications **13**(4), 18–28 (1998)

[14] Hecht-Nielsen, R.: Theory of the backpropagation neural network. In: Neural networks for perception, pp. 65–93. Elsevier (1992)

[15] Hinton, G.: Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531 (2015)

[16] Howard, A., Sandler, M., Chu, G., Chen, L.C., Chen, B., Tan, M., Wang, W., Zhu, Y., Pang, R., Vasudevan, V., et al.: Searching for mobilenetv3. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 1314–1324 (2019)

[17] Howard, A.G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., Adam, H.: Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861 (2017)

[18] Iddan, G., Meron, G., Glukhovsky, A., Swain, P.: Wireless capsule endoscopy. Nature **405**(6785), 417–417 (2000)

[19] Johnson, L.R.: Physiology of the gastrointestinal tract. Elsevier (2006)

[20] Le Floch, M., Wolf, F., McIntyre, L., Weinert, C., Palm, A., Volk, K., Herzog, P., Kirk, S.H., Steinhaeuser, J.L., Stopp, C., Geissler, M.E., Herzog, M., Sulk, S., Kather, J.N., Meining, A., Hann, A., Hampe, J., Herzog, N., Brinkmann, F.: Galar-a large multi-label video capsule endoscopy dataset. medRxiv pp. 2024–09 (2024)

[21] Liao, J., Cai, L., Xu, Y., He, M.: Design of accelerator for mobilenet convolutional neural network based on fpga. In: 2019 IEEE 4th Advanced Information Technology, Electronic and Automation Control Conference (IAEAC). vol. 1, pp. 1392–1396. IEEE (2019)

[22] Mylonaki, M., Fritscher-Ravens, A., Swain, P.: Wireless capsule endoscopy: a comparison with push enteroscopy in patients with gastroscopy and colonoscopy negative gastrointestinal bleeding. Gut **52**(8), 1122–1126 (2003)

[23] Regmi, S., Subedi, A., Bagci, U., Jha, D.: Vision transformer for efficient chest x-ray and gastrointestinal image classification. arXiv preprint arXiv:2304.11529 (2023)

[24] de Sá, D.G., Freulonx, G.d.A., Ferreira, M.P., Pessoa, A.C., Quintanilha, D.B., Silva, A.C.: Abnormality detection in wireless capsule endoscopy images using deep features. In: International Conference on Wireless Mobile Communication and Healthcare. pp. 173–184. Springer (2023)

[25] Sagi, O., Rokach, L.: Ensemble learning: A survey. Wiley interdisciplinary reviews: data mining and knowledge discovery **8**(4), e1249 (2018)

[26] Savio, A., García-Sebastián, M.T., Chyzyk, D., Hernández, C., Graña, M., Sistiaga, A., De Munain, A.L., Villanúa, J.: Neurocognitive disorder detection based on feature vectors extracted from vbm analysis of structural mri. Computers in biology and medicine **41**(8), 600–610 (2011)

[27] Schindler, R., Eusterman, G.B.: Gastroscopy: the endoscopic study of gastric pathology. Annals of Surgery **106**(5), 958 (1937)

[28] Shahzad, F., Mannan, A., Javed, A.R., Almadhor, A.S., Baker, T., Al-Jumeily OBE, D.: Cloud-based multiclass anomaly detection and categorization using ensemble learning. Journal of Cloud Computing **11**(1), 74 (2022)

[29] Smedsrud, P.H., Thambawita, V., Hicks, S.A., Gjestang, H., Nedrejord, O.O., Næss, E., Borgli, H., Jha, D., Berstad, T.J.D., Eskeland, S.L., et al.: Kvasir-capsule, a video capsule endoscopy dataset. Scientific Data **8**(1), 142 (2021)

[30] Srivastava, A., Tomar, N.K., Bagci, U., Jha, D.: Video capsule endoscopy classification using focal modulation guided convolutional neural network. In: 2022 IEEE 35th International Symposium on Computer-Based Medical Systems (CBMS). pp. 323–328. IEEE (2022)

[31] Vanerio, J., Casas, P.: Ensemble-learning approaches for network security and anomaly detection. In: Proceedings of the workshop on big data analytics and machine learning for data communication networks. pp. 1–6 (2017)

[32] Wen, Y., Tran, D., Ba, J.: Batchensemble: an alternative approach to efficient ensemble and lifelong learning. arXiv preprint arXiv:2002.06715 (2020)

[33] Williams, C., Teague, R.: Colonoscopy. Gut **14**(12), 990 (1973)

[34] Yan, S., Liu, Z., Wang, Y., Zeng, C., Liu, Q., Cheng, B., Cheung, R.C.: An fpga-based mobilenet accelerator considering network structure characteristics. In: 2021 31st International Conference on Field-Programmable Logic and Applications (FPL). pp. 17–23. IEEE (2021)

[35] Zhang, Z., Sabuncu, M.: Generalized cross entropy loss for training deep neural networks with noisy labels. Advances in neural information processing systems **31** (2018)

[36] Zhou, C., Paffenroth, R.C.: Anomaly detection with robust deep autoencoders. In: Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining. pp. 665–674 (2017)

[37] Zong, B., Song, Q., Min, M.R., Cheng, W., Lumezanu, C., Cho, D., Chen, H.: Deep autoencoding gaussian mixture model for unsupervised anomaly detection. In: International conference on learning representations (2018)

Seeing More with Less: Video Capsule Endoscopy with Multi-Task Learning

Julia Werner[1][0009-0006-0279-1776], Oliver Bause[1][0009-0003-5388-2959], Julius Oexle[1], Maxime Le Floch[2,3], Franz Brinkmann[2,3][0000-0002-3474-3115], Jochen Hampe[2,3][0000-0002-2421-6127], and Oliver Bringmann[1][0000-0002-1615-507X]

[1] Department of Computer Science, University of Tübingen, Tübingen, Germany
[2] Else Kröner Fresenius Center for Digital Health, TU Dresden, Dresden, Germany
[3] Department of Medicine I, University Hospital Dresden, TU Dresden, Dresden, Germany

Abstract. Video capsule endoscopy has become increasingly important for investigating the small intestine within the gastrointestinal tract. However, a persistent challenge remains the short battery lifetime of such compact sensor edge devices. Integrating artificial intelligence can help overcome this limitation by enabling intelligent real-time decision-making, thereby reducing the energy consumption and prolonging the battery life. However, this remains challenging due to data sparsity and the limited resources of the device restricting the overall model size. In this work, we introduce a multi-task neural network that combines the functionalities of precise self-localization within the gastrointestinal tract with the ability to detect anomalies in the small intestine within a single model. Throughout the development process, we consistently restricted the total number of parameters to ensure the feasibility to deploy such model in a small capsule. We report the first multi-task results using the recently published Galar dataset, integrating established multi-task methods and Viterbi decoding for subsequent time-series analysis. This outperforms current single-task models and represents a significant advance in AI-based approaches in this field. Our model achieves an accuracy of 93.63% on the localization task and an accuracy of 87.48% on the anomaly detection task. The approach requires only 1 million parameters while surpassing the current baselines.

Keywords: Video Capsule Endoscopy · Multi-Task Learning · Viterbi decoding.

1 Introduction

The Video Capsule Endoscopy (VCE) aims to detect pathological tissue within the gastrointestinal (GI) tract, more specifically the small intestine, by employing

a small pill-size capsule, which is equipped among others with LEDs and a camera [8, 20]. By peristalsis, the capsule traverses through the GI tract and captures low-resolution images throughout its journey, which are then sent out to an external device for final evaluation by medical doctors. This procedure is essential as it transmits images of the small intestine, which is largely not accessible by standard techniques, such as the gastroscopy or colonoscopy [5,18]. Thus, the main objective of the VCE is to cover the small intestine, while the esophagus, stomach and colon are of less interest.

Current devices, such as the PillCam SB3 [14], simply transmit all captured images to an external device without prior classification. However, this transmission is very costly and only those frames originating from the small intestine are actually relevant in the VCE. On the other hand, simply storing all images on the device locally until the end of the procedure requires collection of the capsule afterwards [27], does not improve the battery lifetime and prohibits real-time assessment as well as decision-making. Importantly, as the available energy of such small sensor edge device is very limited, e.g. the PillCam has a battery lifetime of 8-12h [14, 15], it is essential to minimize unnecessary energy usage. Since the traversal time of the capsule varies between patients, for some patients, the small intestine is not covered by such capsule, before the battery is depleted. Theoretically, images that are not of interest can be discarded immediately, saving the energy that would be otherwise be used for their transmission. To achieve this, the organ, which the capsule currently traverses, needs to be classified and transmission of images only started after exiting the stomach.

Furthermore, jointly addressing organ detection with anomaly detection while continuously limiting the overall model size, can improve the entire procedure by ensuring comprehensive coverage of the GI tract and enabling preliminary assessments of potential anomalies prior to final evaluation by physicians. On-site anomaly detection allows real-time decisions to be made, such as raising the frame rate or resolution, which might improve the visualization of important regions. However, AI-based anomaly detection still remains a major challenge in this field due to data sparsity and the need of high medical competence to label such data. One possible technique to tackle this is multi-task learning (MTL) [3] which leverages the observation that, in certain cases, simultaneous learning of multiple tasks can produce superior results [3, 26]. This approach relies on the assumption that the tasks being targeted are related. When this condition is met, MTL has the potential to surpass the performance of single-task learning (STL).

Our Contribution: By applying MTL to VCE, our main objective is to develop a model capable of concurrently classifying the organ section and assessing whether an anomaly is present in the captured image. Precisely determining the capsule's entry into the small intestine enables suppression of image transmission outside the target region and thus, reduces energy consumption. This helps to ensure full small intestine coverage across all patients without exhausting the battery. Thus, we report the first multi-task results targeting this localization problem along with anomaly detection and aim to surpass the

performance of current state-of-the-art single-task models for both objectives. Since for real-world applications, this model must be deployed on a very small device, its hardware implementation feasibility is a key consideration. As an initial step, we focus on limiting the overall model size and the total number of multiply-accumulate (MAC) operations compared to the current baselines.

2 Related Work

Multi-Task Studies Targeting the Gastrointestinal Tract: Recently, several MTL approaches have been developed to capture diseases in the GI tract. For example, [10] targeted the detection of Crohn's disease, by using MTL with a private dataset including standard endoscopy images from 15 patients, however, using very large networks, such as a DenseNet121 and a ResNet50, for this task. More research has been conducted to detect esophageal lesions [25] and polyps in the colon [19] with MTL. Notably, both studies did not target the small intestine. Most recently, [23] performed magnetically controlled capsule endoscopy for the classification of gastric anatomical sites as well as lesions on a public, self-built dataset. In this work, following the constraints of VCE applications, only datasets containing low-resolution images from VCE studies with a specific focus on targeting the small intestine are applicable.

Localization and Anomaly Detection with VCE Images: The reported results are predominantly limited to single-task models, which can either identify the anatomical regions or classify lesions in the digestive tract. This is inherently given by the fact that the previously published VCE datasets contain labels only for a single task. For example, the Kvasir-Capsule dataset [18] consists only of images labeled for anomalies but not the organ sections while the Rhode-Island Gastroenterology dataset [4] comprises of images labeled for the different organ sections but not for any anomalies. The newly available large Galar dataset [11] incorporates images labeled not only for different GI sections but also anomalies. Besides [11], [21] were the first to perform anomaly detection on the Galar dataset with an ensemble model, however; without targeting the localization task. We intend to perform MTL on VCE images drawn from this dataset and outperform current single-task baselines by combining the localization and the anomaly detection task while restricting the overall model size. Since the small intestine is the only part of the GI tract largely inaccessible by standard endoscopes, it is the primary focus of this procedure.

3 Experiments and Methodology

The primary goal of our method is to prepare a single, resource-constrained, light-weight neural network, that is capable of performing both, localization and anomaly detection tasks, on VCE images as presented in Figure 1 to ensure coverage of the small intestine before the capsule's battery is depleted and provide basic anomaly detection functionalities. We start by preparing the dataset to

c
l
t

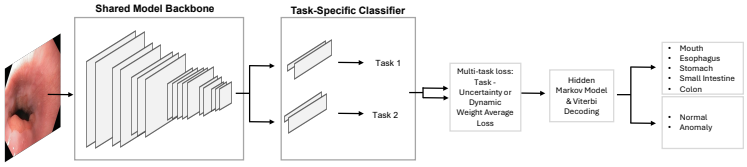


Fig. 1: The neural network is trained on the Galar dataset with VCE images and hard parameter sharing with two separate classifier heads, dynamic weight average or task uncertainty weighting of losses and post-processing with a HMM and Viterbi decoding.

3.1 Pre-Processing - Galar Dataset

To the best of our knowledge, the Galar dataset [11] is the only publicly available VCE dataset, which consists of images labeled for the different GI organs as well as anomalies. Thus, this dataset is well-suited for simulating the VCE environment and evaluating multi-task learning approaches for this application. The authors of the Galar dataset selected patients with the ID $61 - 80$ for the test set, representing completely untouched VCE studies from a clinical setting. For better comparability, we evaluate our method on the same test set. The remaining 60 patient studies (ID $1 - 60$) were used as a development set for training the neural network.

As multiple other medical datasets, the Galar dataset is characterized by a large class imbalance. From all organs available in this dataset, the smallest class (mouth) represents only 0.05 % of all images. Moreover, pathological cases represent the minority within the classes, and in addition, there are significant differences in class frequencies among the pathologies themselves. While the smallest anomaly class (erythema) comprises only 0.12 % of the data, the largest class (blood) makes up 11.9 % of all frames. Additionally, if the capsule traveled slowly or resides at one location for some time, many similar images are often captured, which can disrupt the training process. To counteract this class imbalance, the majority classes were downsampled as described in the following.

The main objective remains to theoretically capture all images of interest and transmit them to an on-body device for final evaluation by physicians. We do not aim to provide an extensive evaluation of anomalies on-site, but to capture optimally all relevant images, which are finally evaluated by medical doctors after transmission. Additionally, basic on-site anomaly detection allows important

features to be added, such as increased resolution or frame rate upon outlier detection improving the entire procedure. Thus, detailed analysis of different anomaly types does not need to be conducted on the capsule directly. Instead, using a binary distinction eases the computational workload of the model for this application and also enables size restrictions to be applied. Therefore, all pathologies were combined into a single anomaly class in order to perform binary classification. For each pathology, a random selection of frames was included. Next, all samples from the mouth and esophagus were collected and the remaining section classes randomly downsampled to have a final ratio of $1:1$ (normal : anomaly). The data was randomly shuffled and split into a training and validation set with a defined ratio of $70:30$. The final class distribution of the different organ sections and the anomaly and normal samples is depicted in Table 1.

Table 1: Class distribution of the used organ sections and anomalies (positive samples) after downsampling to address the occurrence of many related frames.

Subset	Mouth	Esophagus	Stomach	Small Intestine	Colon	Total	Negative	Positive	Total
Train	1,476	1,718	11,962	32,274	26,059	73,489	39,779	40,104	79,883
Val	252	132	2,598	9,478	7,633	20,093	10,748	10,757	21,505
Test	265	397	16,510	175,716	41,667	234,555	218,665	16,499	235,164

3.2 Neural Network Training with Hard-Parameter Sharing

In the context of capsule endoscopy, multiple authors [22, 23] aimed to restrict the total number of parameters to address the constraints given by tiny edge devices equipped during such procedure by using light-weight networks from the MobileNet family. These networks are not only beneficial in image classification tasks but were additionally designed for the efficient deployment on embedded devices [7, 17]. Hence, in this work, we adopt the MobileNetV3-Small [7] architecture and modify the classification head. Instead of a single classification head, we designed and generated separate classification heads tailored to each specific task. Offering the benefit of minimizing parameters and straightforward deployment while allowing the classification of multiple tasks, we implement hard parameter sharing [2]. This MTL technique is used to share the hidden layers between all tasks and further exploit task-specific classifier heads.

The pretrained model was fine-tuned for additional 10 epochs and the learning rate and weight decay were each defined based on a hyperparameter search with Hydra [24] and the well-established hyperparameter optimization framework Optuna [1] with 30 trials (for the localization, the anomaly detection and the multi-task runs each). We aimed to employ well-established MTL losses for

comparison and to investigate how this influences the final results. Thus, the homoscedastic task uncertainty based on [9] was implemented. The MTL loss $\mathcal{L}$ is assembled for the tasks $i \in \{1, 2\}$, with

$$\mathcal{L} = \sum_{i=1}^{n} \exp\left(-\log \sigma_i\right) \mathcal{L}_i + \log \sigma_i. \tag{1}$$

Furthermore, the Dynamic Weight Average Loss (DWA) [13] was implemented to improve the balancing of the loss values for both tasks. Instead of weighting the tasks according to their homoscedastic uncertainty, this loss adjusts its weights based on loss changes between the epochs. This helps to prevent premature neglect of any task and supports the learning of more challenging tasks over extended training periods. The temperature factor T also allows the strength of the weighting to be set manually. The DWA weighting λ_i is assembled for the tasks $i \in \{1, 2\}$, with

$$\lambda_i(t) := \frac{I \exp\left(w_i(t-1)/T\right)}{\sum_k \exp\left(w_k(t-1)/T\right)}, w_i(t-1) = \frac{\mathcal{L}_i(t-1)}{\mathcal{L}_i(t-2)}. \tag{2}$$

For the anomaly detection task, the cross-entropy loss was replaced with a focal loss [12] to address the pronounced class imbalance between normal and abnormal tissue samples. This allows the model to focus more on the underrepresented class without requiring oversampling of abnormal cases, thereby preserving a realistic class distribution during training. The final evaluations were then passed to a HMM to perform post-processing.

3.3 Post-Processing - Hidden Markov Model and Viterbi Decoding

The HMM [16] has previously been applied to the problem of localization in video capsule endoscopy, using the only available dataset at the time with labels for 4 organ sections: the Rhode Island VCE dataset [4] [22]. It has been shown, that specifically for this setting involving time-series data, post-processing the CNN output with a HMM and Viterbi decoding leads to a significant improvement in performance allowing a more precise determination of when the small intestine is entered compared to solely using a CNN. This method is particularly well-suited to the problem at hand as it incorporates prior knowledge of the sequential order in which the organs are traversed and their respective lengths.

In this work, we aim to transfer this technique to the newly available Galar dataset that includes class labels for the organs across the GI tract and apply it as a post-processing to the MTL approach. Compared to [22], the number of hidden states is extended to 5, such that we have $\{S_1 = \text{Mouth}, S_2 = \text{Esophagus}, S_3 = \text{Stomach}, S_4 = \text{Small intestine}, S_5 = \text{Colon}\}$, with an initial distribution $\pi_i = P(X_1 = S_i)$ for $i \leq 5$.The HMM for this GI setting is visualized in Figure 2.

Each moment t, at which the capsule captured an image, an observation $X_t \in K_1, \ldots, K_m$ is obtained, representing a location classified by the neural network. The transition probabilities $a_{ij} = P(X_{t+1} = S_j | X_t = S_i)$ of the

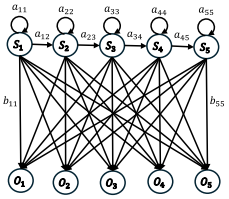


Fig. 2: Hidden Markov Model for the described GI setting.

Markov chain S_t are encoded such that for S_{t+1} only the current organ or the succeeding organ can be the next state, in accordance with the natural anatomy of the GI tract. The emission probabilities $b_j(k) = P(O_t = K_k | X_t = S_j)$ for an observation $O_t \in \{O_1 = \text{Mouth}, O_2 = \text{Esophagus}, O_3 = \text{Stomach}, O_4 = \text{Small intestine}, O_5 = \text{Colon}\}$ at time t are inferred based on the confusion matrix from the neural network on the training set as this directly encodes the emission probabilities. Finally, the Viterbi algorithm [6] calculates the most probable sequence of gastroenterological states $X = (X_1, \ldots, X_t)$, based on the evaluations of the neural network $(O_1, \ldots, O_t)$, with $\arg\max_{X_1,\ldots,X_t} P(X_1, \ldots, X_t, O_1, ...O_t)$ as described in [22].

4 Results

Our main objective is to provide a very light-weight multi-task model, that enhances the current baselines of single-task models, while reducing the overall complexity and provide useful functionalities for VCE devices, predominantly by ensuring the coverage of the whole small intestine by precise organ detection/localization of the capsule. Typically, deep neural networks are too large to be deployed on tiny edge devices, such as the video capsule for endoscopies. To address this, we conducted all experiments using the light-weight MobileNet [7]. First, in Table 2, the results for the organ detection are presented.

The ST model shows inferior results compared to the baseline with a F1-score of only 60.55% possibly due to a less complex model, however, this is notably improved by the MT approach. The proposed MT approach, if combined with Viterbi decoding, shows a strong performance across all tested loss functions. The best results were achieved with the DWA and focal loss, with an accuracy of 93.63% and a F1-score of 92.41 compared to an accuracy of 81% and F1-score of 71% of the baseline [11]. We further observe, that the uncertainty weighted loss shows inferior results compared to the DWA or DWA with focal loss.

Furthermore, in Table 3, the MT results for the anomaly detection task are shown. While all three loss options lead to proficient results, the DWA loss produces the best results compared to the baseline and the ST model (F1-score: 54.38% vs. 37.01%). Only the recall is inferior compared to the baseline, with only 54.69% vs. 60.88%. However, this work provides a VCE targeted approach

Table 2: Results of the multi-task (MT) runs for the **localization** task in [%] with the **MobileNetV3** compared to single-task (ST) baseline results with the best results in bold and the second best results underlined.

	Accuracy	F1	Precision	Recall	Params	MACs
Baseline [11]	81	71	–	–	25 M	4,087 B
Our ST Localization	85.05	60.55	54.42	86.44	1 M	63,614 M
Our MT Results						
MT UW w/o HMM	80.96	49.30	48.50	69.58	1 M	63,616 M
MT DWA w/o HMM	86.22	61.53	54.48	87.08	1 M	63,616 M
MT DWA Focal w/o HMM	87.77	65.12	58.61	87.5	1 M	63,616 M
MT UW & HMM	90.48	66.22	71.57	79.14	1 M	63,616 M
MT DWA & HMM	<u>93.48</u>	<u>91.02</u>	<u>88.56</u>	<u>94.34</u>	1 M	63,616 M
MT DWA Focal & HMM	**93.63**	**92.41**	**90.40**	**94.94**	1 M	63,616 M

to perform the important tasks of localization and anomaly detection within one single model. Notably, both tasks can be conducted with the same model of only 1 M parameters, which is only 25% of the anomaly detection baseline model and merely 4% of the localization baseline model. In addition, the number of MAC units is reduced from 4 B and 189 M to only 64 M with this approach, lowering the computational workload.

Table 3: Results of the multi-task (MT) runs for the **anomaly detection** task in [%] with the **MobileNetV3** compared to single-task (ST) baseline results.

Classification	Accuracy	F1	Precision	Recall	Params	MACs
Baseline [21]	87.28	37.01	26.59	**60.88**	4 M	189 M
Our ST Anomaly Detection	84.91	<u>53.34</u>	52.99	54.63	1 M	63,611 M
Our MT Results						
MT UW	83.69	52.94	52.71	54.66	1 M	63,616 M
MT DWA	**87.7**	**54.38**	**54.14**	<u>54.69</u>	1 M	63,616 M
MT DWA Focal	<u>87.48</u>	53.24	<u>53.08</u>	53.45	1 M	63,616 M

5 Conclusion

This work introduces a light-weight neural network with a total of just 1 M parameters, designed to combine precise self-localization and anomaly detection capabilities through a multi-task learning approach for VCE. We report the first multi-task results using the recently published Galar dataset and integrate various established multi-task methods and Viterbi decoding. With an accuracy of 93.63 % and an F1-score of 92.41 %, our approach shows superior performance

for the localization task compared to the current baseline, while also enabling fundamental anomaly detection functionality. Thus, the presented work improves upon current AI models for VCE, by providing precise localization within the GI tract and basic anomaly detection functionalities.

Prospect of application: Once the exit of the stomach has been determined by the multi-task model, additional features such as adapting frame rates (lowered before the small intestine and increased upon anomaly detection) or varying resolutions may be added. The approach can prolong battery life and serve as a starting point for deploying multi-task models on VCE devices.

Acknowledgments This work has been partly funded by the German Federal Ministry of Research, Technology and Space (BMFTR) in the project MEDGE (16ME0530).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Akiba, T., Sano, S., Yanase, T., Ohta, T., Koyama, M.: Optuna: A next-generation hyperparameter optimization framework. In: Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining. pp. 2623–2631 (2019)
2. Caruana, R.: Multitask learning: A knowledge-based source of inductive bias1. In: Proceedings of the Tenth International Conference on Machine Learning. pp. 41–48. Citeseer (1993)
3. Caruana, R.: Multitask learning. Machine learning **28**, 41–75 (1997)
4. Charoen, A., Guo, A., Fangsaard, P., Taweechainaruemitr, S., Wiwatwattana, N., Charoenpong, T., Rich, H.G.: Rhode island gastroenterology video capsule endoscopy data set. Scientific Data **9**(1), 602 (2022)
5. Costamagna, G., Shah, S.K., Riccioni, M.E., Foschia, F., Mutignani, M., Perri, V., Vecchioli, A., Brizi, M.G., Picciocchi, A., Marano, P.: A prospective trial comparing small bowel radiographs and video capsule endoscopy for suspected small bowel disease. Gastroenterology **123**(4), 999–1005 (2002)
6. Forney, G.D.: The viterbi algorithm. Proceedings of the IEEE **61**(3), 268–278 (1973)
7. Howard, A., Sandler, M., Chu, G., Chen, L.C., Chen, B., Tan, M., Wang, W., Zhu, Y., Pang, R., Vasudevan, V., et al.: Searching for mobilenetv3. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 1314–1324 (2019)
8. Iddan, G., Meron, G., Glukhovsky, A., Swain, P.: Wireless capsule endoscopy. Nature **405**(6785), 417–417 (2000)
9. Kendall, A., Gal, Y., Cipolla, R.: Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7482–7491 (2018)
10. Kong, Z., He, M., Luo, Q., Huang, X., Wei, P., Cheng, Y., Chen, L., Liang, Y., Lu, Y., Li, X., et al.: Multi-task classification and segmentation for explicable capsule endoscopy diagnostics. Frontiers in Molecular Biosciences **8**, 614277 (2021)

11. Le Floch, M., Wolf, F., McIntyre, L., Weinert, C., Palm, A., Volk, K., Herzog, P., Kirk, S.H., Steinhäuser, J.L., Stopp, C., et al.: Galar-a large multi-label video capsule endoscopy dataset. Scientific Data **12**(1), 828 (2025)
12. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE international conference on computer vision. pp. 2980–2988 (2017)
13. Liu, S., Johns, E., Davison, A.J.: End-to-end multi-task learning with attention. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1871–1880 (2019)
14. Medtronic: PillCam™ SB3 System. https://www.medtronic.com/covidien/en-nz/products/capsule-endoscopy/pillcam-sb-3-system.html/ (2025), [Online; accessed 7-May-2025]
15. Monteiro, S., de Castro, F.D., Carvalho, P.B., Moreira, M.J., Rosa, B., Cotter, J.: Pillcam® sb3 capsule: Does the increased frame rate eliminate the risk of missing lesions? World journal of gastroenterology **22**(10), 3066 (2016)
16. Rabiner, L.R.: A tutorial on hidden markov models and selected applications in speech recognition. Proceedings of the IEEE **77**(2), 257–286 (1989)
17. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.C.: Mobilenetv2: Inverted residuals and linear bottlenecks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4510–4520 (2018)
18. Smedsrud, P.H., Thambawita, V., Hicks, S.A., Gjestang, H., Nedrejord, O.O., Næss, E., Borgli, H., Jha, D., Berstad, T.J.D., Eskeland, S.L., et al.: Kvasir-capsule, a video capsule endoscopy dataset. Scientific Data **8**(1), 142 (2021)
19. Tang, S., Yu, X., Cheang, C.F., Liang, Y., Zhao, P., Yu, H.H., Choi, I.C.: Transformer-based multi-task learning for classification and segmentation of gastrointestinal tract endoscopic images. Computers in Biology and Medicine **157**, 106723 (2023)
20. Thomson, A., Keelan, M., Thiesen, A., Clandinin, M., Ropeleski, M., Wild, G.: Small bowel review: diseases of the small intestine. Digestive diseases and sciences **46**, 2555–2566 (2001)
21. Werner, J., Gerum, C., Nick, J., Floch, M.L., Brinkmann, F., Hampe, J., Bringmann, O.: Enhanced anomaly detection for capsule endoscopy using ensemble learning strategies. arXiv preprint arXiv:2504.06039 (2025)
22. Werner, J., Gerum, C., Reiber, M., Nick, J., Bringmann, O.: Precise localization within the gi tract by combining classification of cnns and time-series analysis of hmms. In: International Workshop on Machine Learning in Medical Imaging. pp. 174–183. Springer (2023)
23. Xu, T., Li, Y.Y., Huang, F., Gao, M., Cai, C., He, S., Wu, Z.X.: A multi-task neural network for image recognition in magnetically controlled capsule endoscopy. Digestive Diseases and Sciences pp. 1–9 (2024)
24. Yadan, O.: Hydra - a framework for elegantly configuring complex applications. Github (2019), https://github.com/facebookresearch/hydra
25. Yu, X., Tang, S., Cheang, C.F., Yu, H.H., Choi, I.C.: Multi-task model for esophageal lesion analysis using endoscopic images: classification with image retrieval and segmentation with attention. Sensors **22**(1), 283 (2021)
26. Zhang, Y., Yang, Q.: A survey on multi-task learning. IEEE transactions on knowledge and data engineering **34**(12), 5586–5609 (2021)
27. Zwinger, L.L., Siegmund, B., Stroux, A., Adler, A., Veltzke-Schlieker, W., Wentrup, R., Jürgensen, C., Wiedenmann, B., Wiedbrauck, F., Hollerbach, S., et al.: Capsocam sv-1 versus pillcam sb 3 in the detection of obscure gastrointestinal bleeding:

results of a prospective randomized comparative multicenter study. Journal of clinical gastroenterology **53**(3), e101–e106 (2019)

Reliable Mislabel Detection for Video Capsule Endoscopy Data

1st Julia Werner
Embedded Systems
University of Tübingen
Tübingen, Germany

2nd Julius Oexle
Embedded Systems
University of Tübingen
Tübingen, Germany

3rd Oliver Bause
Embedded Systems
University of Tübingen
Tübingen, Germany

4th Maxime Le Floch
Department of Medicine I
University Hospital Dresden
TU Dresden, Germany

5th Franz Brinkmann
Department of Medicine I
University Hospital Dresden
TU Dresden, Germany

6th Hannah Tolle
Department of Medicine I
University Hospital Dresden
TU Dresden, Germany

7th Jochen Hampe
Department of Medicine I
University Hospital Dresden
TU Dresden, Germany

8th Oliver Bringmann
Embedded Systems
University of Tübingen
Tübingen, Germany

Abstract—The classification performance of deep neural networks relies strongly on access to large, accurately annotated datasets. In medical imaging, however, obtaining such datasets is particularly challenging since annotations must be provided by specialized physicians, which severely limits the pool of annotators. Furthermore, class boundaries can often be ambiguous or difficult to define which further complicates machine learning-based classification. In this paper, we want to address this problem and introduce a framework for mislabel detection in medical datasets. This is validated on the two largest, publicly available datasets for Video Capsule Endoscopy, an important imaging procedure for examining the gastrointestinal tract based on a video stream of low-resolution images. In addition, potentially mislabeled samples identified by our pipeline were reviewed and re-annotated by three experienced gastroenterologists. Our results show that the proposed framework successfully detects incorrectly labeled data and results in an improved anomaly detection performance after cleaning the datasets compared to current baselines.

Index Terms—Video Capsule Endoscopy, Unsupervised Noise Detection, Anomaly Detection, Dataset Cleaning

I. Introduction

Proficient performance of machine learning models highly depends on access to large, representative datasets. However, when large-scale datasets are annotated by humans, the occurence of mislabeled samples is inevitable and a general assumption that all annotations are accurate can introduce substantial issues. Since deep neural networks tend to overfit on noisy labels, such label noise can adversely affect the generalization and prediction performance [24], [28]. In medical applications, the annotation of real-world clinical data is particularly challenging, as it is time-consuming and typically requires specialized clinical expertise, which restricts the pool of qualified annotators.

An example for such medical application is the Video Capsule Endoscopy (VCE), a key diagnostic medical procedure to examine the gastrointestinal (GI) tract, that was first introduced in the early 2000s [8], [27], [29]. The VCE is specifically applied to inspect the small intestine for pathologies while the stomach and colon can be investigated by standard procedures such as a gastroscopy or colonoscopy [3], [26]. This procedures involves a small pill-sized capsule consisting of a camera, a transmitter, a battery and LEDs, that can be swallowed by patients to record the inside of the digestive tract while it moves through the gastrointestinal organs (mouth, esophagus, stomach, small intestine, colon) [17], [18]. With current devices on the market, images recorded by the camera are directly transmitted to an on-body receiver for subsequent assessment by gastroenterologists [17]. For this application, the long-term objective is on-device anomaly detection in real-time to enable timely diagnosis.

To realize successful screening for pathologies, vision models can be employed, which, however, require access to adequate datasets for effective neural network training. Importantly, VCE datasets are characterized by a large imbalance with the number of anomalies in the minority [11], [26], [32]. In this case, mislabeled data has an even stronger impact and incorrect labels can negatively influence the results strongly. The general presence of low-resolution images obtained in VCEs further complicate evaluations. Additionally, fine-grained annotations can be subjective, as the transition from one class to another can be blurred. To circumvent the potential label ambiguity, one can either increase the robustness of machine learning models or screen for noisy samples and process them prior to neural network training [10], [21], [34].

Our Contribution: In this work, we developed a framework for mislabel detection in endoscopy datasets and evaluate its performance on the largest, publicly available VCE datasets: Kvasir-Capsule [26] and Galar [11]. To further assess its reliability, a subset of samples underwent internal review by three co-authors who are experienced

This work has been partly funded by the German Federal Ministry of Research, Technology and Space (BMFTR) in the project MEDGE (16ME0530).

gastroenterologists. Two of these authors were also among original authors of the relevant VCE dataset. The review was conducted as part of the collaborative authorship process and served as an expert panel. Incorporating this noise detection pipeline into an anomaly detection workflow leads to substantial improvement compared to the usage of uncleaned datasets and existing baselines. Accordingly, this work proposes an approach for processing medical image datasets before neural network training to boost the overall classification performance.

II. Related Work

There are various strategies to handle mislabeled data in the field of machine learning, which focus on increasing the robustness of the models to noisy data during the training process [21]. For instance, instead of adjusting the data points directly, loss functions specifically designed to be less sensitive to erroneous gradients induced by mislabeled samples can be employed [21], [30], [35]. Additionally, there are methods to clean or filter existing datasets from potential mislabels, e.g. by using ensemble methods [5], [20], or to correct noisy data [10], [34]. Some approaches for reducing label noise are based on the assumption that samples close to each other in a feature space are similar and will therefore have the same label. Meanwhile, deviating samples with the same label might be considered noisy [12], [25]. Confident learning has also been used to determine mislabels by exploiting the confidence of a neural network during training. This technique is based on the assumption that reduced prediction confidence is indicative of mislabeled samples [19]. Nevertheless, confidence must be contextualized by the difficulty of classifying each individual sample.

To provide a machine learning-based anomaly detection model for the VCE, large and high quality VCE datasets are essential but remain sparse as they are difficult to be produced and annotated. Recently, the Galar dataset was published, as the largest multi-label VCE datasets [11]. Notably, even with such an extensive and high quality dataset, anomaly detection remains difficult for this application. For example, [11] achieve a F1-score 5% for polyp detection and 14% for blood detection. While this was improved in later publications to a F1-score of 37.01% [32] and 54.38% [31] for multiple pathologies pooled together, it emphasizes that there is still room for improvement. Rather than proposing a new anomaly detection model, this work addresses the underlying issue by identifying noisy data and filtering or correcting mislabeled data points, thereby improving the overall neural network performance.

III. Methodology

One main difficulty when aiming to identify mislabels is a missing ground truth in most datasets. To address this, the experimental setup of the presented work follows a two-stage design, as illustrated in Figure 1. First, a controlled experiment on the Kvasir-Capsule dataset [26] is performed, introducing random label noise (1%, 5%, 10%, 15%, 20%) through label flipping. Since the falsely labeled data points are known in this experiment, this allows adequate evaluation of the pipeline.

In the second stage, the designed mislabel detection pipeline is applied to the Galar dataset [11] with the objective to find pre-existing false labels, resulting in a cleaned version of the dataset. For validation, 100 frames identified by our pipeline as potential mislabels were reviewed by a scientific panel of three gastroenterologists and an updated annotation provided. Finally, a CNN is trained on the cleaned dataset and the anomaly detection performance using the original, unaltered test set evaluated to be comparable to existing baselines.

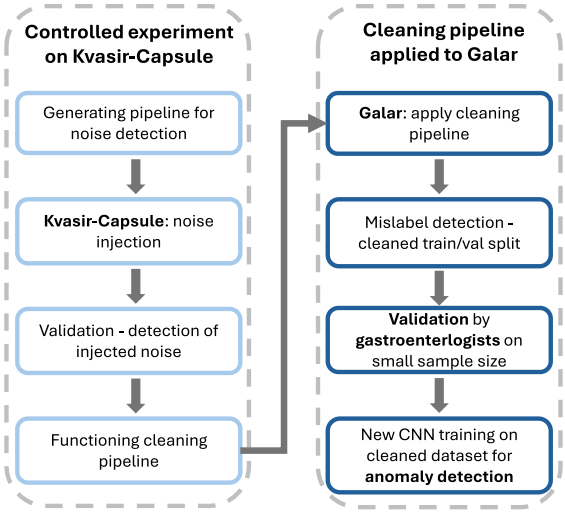


Fig. 1: Experimental design: 1) controlled experiment on the Kvasir-Capsule dataset to identify injected noise and 2) mislabel detection on the Galar dataset involving verification by scientific panel of three gastroenterologists with subsequent anomaly detection.

A. VCE Datasets

The mislabel detection pipeline was validated on the two largest, publicly available VCE datasets to assess its suitability for wireless capsule endoscopy applications. The Kvasir-Capsule dataset [26] contains a total of $47,238$ labeled and $4,694,266$ unlabeled frames. Consistent with the original publication, the official dataset splits were used to guarantee that images from any given patient were not simultaneously present in the development as well as the test set.

Furthermore, the Galar dataset [11] was used with the splits generated as in [31], [32], which comprises a total of $3,513,539$ labeled images. This multi-label VCE dataset consists of annotated frames with anatomical as well as

pathological classes and is intended to be cleaned from noisy data samples in this work.

B. Cleaning Pipeline

Gaussian Mixture Models (GMMs) have been proven to be a useful tool for the detection of noisy samples [13]. As they can obtain an increased loss for mislabeled data points compared to correctly labeled samples, they can be used for noise detection in datasets. Hence, they are an essential component in our experimental setup to detect mislabels as visualized in Figure 2.

First, a neural network is trained on the raw dataset in three different processes with subsequent GMM training. Based on the prediction and confidence of the neural network, for each sample the probability of having a correct label is determined. This value p_i^c describes the noise probability after a potential correction step for a sample with a noise probability p_i. The noise reduction r_i is determined by $p_i - p_i^c$. Then, the first k^c labels presenting the highest noise reduction are corrected (in a binary case, the labels are flipped). Next, three additional CNN and GMM trainings are conducted, concluded with a filtering step in which the first k^f labels with the highest noise probability are filtered to clean the datasets from presumably noisy samples. To determine whether samples with an increased noise probability should be corrected or filtered out, the concept of [9] was applied and the idea of fusing a correction with a filtering step adapted. In the following, the individual components of the cleaning pipeline are explained in more detail.

C. Noise Injection (Kvasir-Capsule dataset)

Noise was injected exclusively for the Kvasir-Capsule dataset, for which the true labels are known and serve as ground truth during evaluation. For each sample, the average prediction confidence and entropy were computed across all epochs and three independent training runs, normalized, equally weighted, and combined into a single uncertainty score reflecting classification difficulty. Based on this score, samples were assigned to low-, mid-, and high-uncertainty quantiles. To imitate noisy samples, label noise was then introduced primarily by randomly selecting samples from the mid- and high-uncertainty groups. To preserve the original class distribution, labels were flipped proportionally within each class.

D. CNN Training

To counteract the influence of the present class imbalances, a focal loss [14] was applied during neural network training. Given that a low model complexity is a key requirement when targeting embedded devices, a MobileNetV3 [6] was used, which combine the suitability for low-power embedded devices with a reliable classification performance [1], [7]. Following [33] who used this model type to classify VCE images while considering the total model complexity of 1 M parameters, this network was employed for VCE image classification.

For neural network training, the HANNAH framework [2] was employed using a learning rate and weight decay of 1×10^{-4} for regularization. The models were trained for 15 epochs if the Galar dataset was used and for 10 epochs if the significantly smaller Kvasir-Capsule dataset was employed, generally using the AdamW optimizer [15].

E. Gaussian Mixture Model

A Gaussian Mixture Model (GMM) is a probabilistic model which assumes that all data points of a mixture were generated from a finite number of gaussian distributions [16], [23]. Each component is represented by a normal distribution, which is defined by a mean μ_k and a variance σ_k^2. Additionally, each component contains a mixture coefficient π_k, describing the number of data points belonging to this component. The GMM can be trained using the Expectation-Maximization-Algorithm (EM) [4], which consists of an expectation and a maximization step.

Using GMMs, we can identify altered losses for noisy samples compared to correctly labeled samples. Therefore, within this work, the sklearn implementation of a three-component GMM was used [22] and trained with the average loss values of each sample per epoch and process. Subsequently, for each sample it was predicted with which probability it belongs to the gaussian distribution with the highest mean. This probability determines the present noise probability p_i of a sample i. The component with the highest mean was determined as the distribution of the wrongly labeled samples and used to compute the noise probability. The component with the lowest mean was determined as the distribution of the samples with correct labels and the third in between those was used as a distribution for difficult samples to learn.

F. Clinical Validation (Galar dataset)

For the clinical validation of the detected label noise in the Galar dataset, a subset of 100 samples in total was selected for subsequent, internal review by three experienced gastroenterologists who are co-authors 4, 5, and 6 of this manuscript. Among them, co-authors 4 and 6 were also involved in the original creation of the Galar dataset. For the subset selection, all samples were ranked in descending order according to their noise-reduction score and the top 500 images considered for further selection since these images represented the samples with the highest likelihood of being mislabeled. Finally, a subset of 100 images was chosen, which comprised 70 images depicting normal mucosa and 30 images showing pathological findings to reflect the predominance of normal findings in VCE studies. Since a wireless endoscopic capsule may remain at the same position for several minutes without movement, it can sometimes produce visually indistinguishable images. To ensure a broad coverage as well as variability, images were drawn from more than 50 distinct videos/patients, with a maximum of three samples per video and a minimum

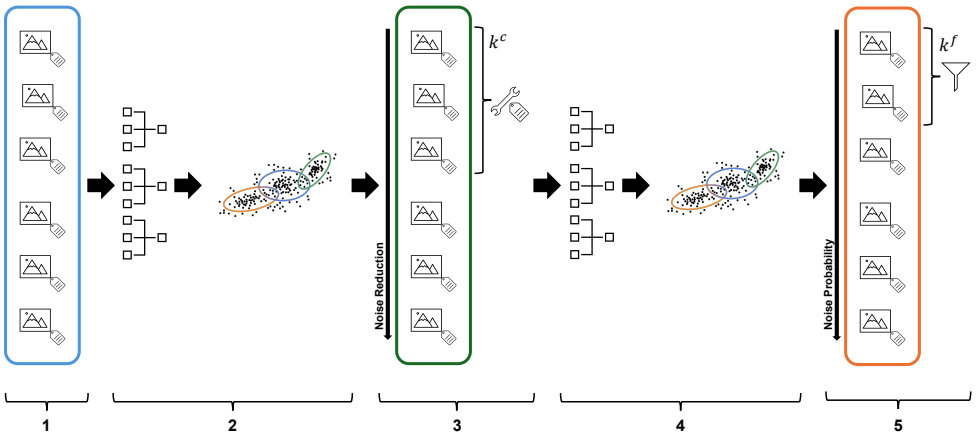


Fig. 2: Mislabel correction pipeline: 1. uncleaned dataset, 2. three CNN trainings with subsequent GMM training, 3. correction of k^c labels based on noise reduction, 4. new training to assess the noise probability, 5. filtering k^f mislabeled data samples.

separation of 100 frames between samples. This constraint mitigates redundancy arising from temporally adjacent frames that may capture nearly identical visual content.

IV. Results

In the following, the results on the Kvasir-Capsule dataset are presented followed by the Galar dataset and finally, the clinical validation by the physicians. First, using the Kvasir-Capsule dataset, the loss density of the different GMM components is inspected to investigate the differences of the three different clusters within the data samples. Figure 3 displays the modeling of the GMM during the correction step with 5% injected noise on the Kvasir-Capsule dataset. This illustrates that the largest number of samples corresponding to correct labels, produces the lowest loss values and is situated below the curve of the first component. The curve of the third component is very flat and contains the highest loss values with strong outliers. For these, it is assumed that the labels are noisy and potentially mislabeled. This illustrates the prominent differences within the used data and this knowledge can be used for mislabel detection.

For the Kvasir-Capsule dataset, for a defined amount of samples (1%, 5% and 10%), the labels are intentionally flipped randomly and afterwards, the cleaning pipeline used to detect the false/noisy labels. Due to this controlled noise injection, after the cleaning pipeline is applied, it can be evaluated proficiently.

Table I presents the mislabel detection results for this experiment and the different amounts of injected noise.

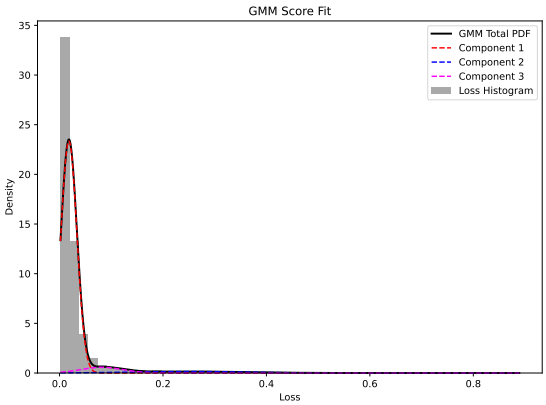


Fig. 3: Distribution of the loss values with the GMM total distribution and the individual components (Kvasir-Capsule dataset, 5% noise injection, correction step).

This shows that the majority of mislabeled samples were correctly detected (e.g. 456/471 for 1% injected noise or 2262/2360 for 5% injected noise).

For the Galar dataset, it is first inspected if images of both classes differ in their structure and how well they can be separated. Figure 4 displays t-Distributed Stochastic Neighbor Embedding (t-SNE) plots on the latent space of the samples in the test set in combination with a Principal

TABLE I: Cleaning status of samples after injecting 1%, 5% or 10% noise to the Kvasir-Capsule dataset.

Cleaning Status (total number of frames)	Injected Noise 1%	5%	10%
Amount of noisy samples	471	2360	4722
Amount of not corrected or filtered samples	15	98	367
Amount of filtered non-noisy samples	916	975	991
Amount of corrected and/or filtered samples	456	2262	4355

Component Analysis (PCA) which reduces the data points to 50 dimensions in total.

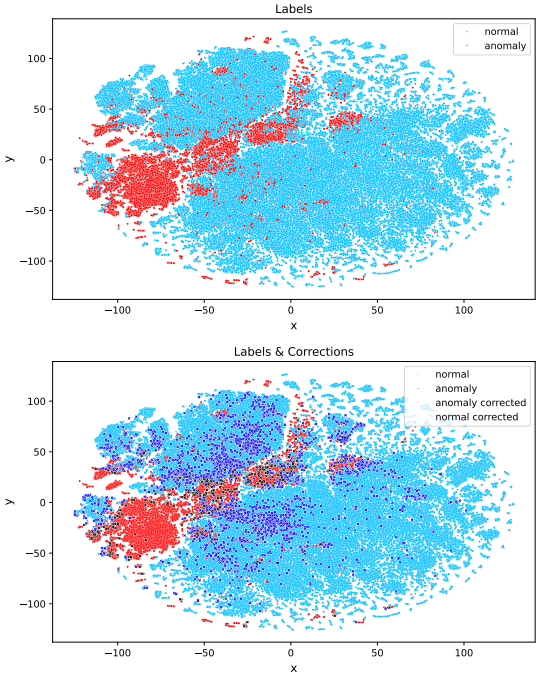


Fig. 4: tSNE visualization of the latent representations before (first plot) and after (second plot) mislabel detection with corrected samples indicated (dark blue: anomaly → normal, black: normal → anomaly).

In the first plot, the samples are colored according to their original annotation in the Galar dataset. A distinct cluster of anomalous samples (red) is visible within the vast majority of normal samples (blue). In addition to this main cluster, individual samples as well as smaller sub-clusters of anomalies occur sporadically within the normal data distribution. The second plot displays the same latent representations but further visualizes the label corrections conducted by our framework. Dark blue corresponds to samples originally annotated as anomalies, but suspected by our methodologies to represent healthy data. Conversely, black indicates samples which are annotated as normal in the Galar dataset, but our framework classified as most presumably being anomalies. This demonstrates that our framework primarily corrects labels whose latent representation lie within the cluster of the other class or in transitional regions between both classes. Consequently, this results in better separated and more coherent clusters. In total, the cleaning pipeline filtered 167,709 samples (4.8% of the Galar dataset) and corrected 31,650 (0.9%) samples. The csv files containing the dataset splits with the proposed correction or filtering applied as well as a file labeling each sample in the entire Galar dataset as anomalous or healthy, as determined by our annotation pipeline, are available here for further usage.

Since a clearer separation of clusters generally facilitates an improved neural network training and leads to an enhanced classification performance, we subsequently evaluated the classification results on the filtered and cleaned dataset compared to the original, uncleaned dataset and the current baselines. While conducting the experiments, the cleaned developmental set was used while the test set was kept untouched for better comparability with the existing baselines. However, in the final run, the test set containing the corrected samples was additionally filtered to investigate if this leads to an additional classification enhancement. The final results are presented in Table II.

TABLE II: Anomaly detection results on the Galar dataset before and after dataset cleaning compared to existing baselines.

Cleaned dev set	Cleaned test set	Accuracy	F1-Score	Precision	Sensitivity	∅ max. confidence
Uncleaned [32]	Uncleaned	87.28	37.01	26.59	60.88	–
Uncleaned [31]	Uncleaned	87.7	54.38	54.14	54.69	–
This work						
Uncleaned	Uncleaned	90.99	53.70	54.69	53.32	0.85
Corrected	Uncleaned	91.48	64.23	63.80	64.69	0.89
Filtered	Uncleaned	93.83	71.58	73.03	70.34	0.96
Filtered	Filtered	91.72	73.67	89.88	68.05	0.96

It is demonstrated that training a neural network on the corrected/filtered training set leads to substantial classification improvement with a total accuracy of 93.83% and a F1-score of 71.58% compared to uncleaned data with an accuracy of 90.99% and a F1-score of 53.70%. Furthermore, a strong advancement can be noted compared to current baselines, which denote F1-scores of 37.01% [32] and 54.38% [31]. It is further observed that the model exhibits a boosted confidence if it was trained on the cleaned dataset vs. the uncleaned dataset. Finally, the precision (89.88%) is strongly enhanced if the test set is also filtered in addition to the dev set. This emphasizes the benefit of using the presented cleaning pipeline before conducting anomaly detection.

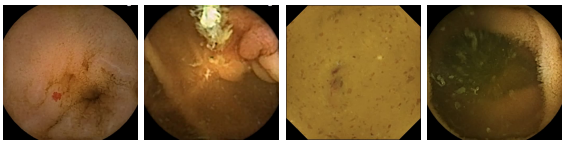
Clinical Validation

Finally, 100 data samples for which our pipeline detected

the highest probability of being mislabeled were re-evaluated by a scientific panel of three clinicians. If at least two physicians determined that a label is wrong, this is considered to be the truth in the following. In total, we received a Precision@100 score of 78. The evaluation is listed in more detailed in the following:

- 78% of suspected labels identified by our pipeline had indeed incorrect labels.
- 49% were originally labeled as normal, but were identified by our pipeline as mislabels and confirmed to be anomalies by the scientific panel.
- 29% were originally labeled as anomaly but were identified by our pipeline as mislabels and confirmed to be indeed normal samples by clinicians.
- Only 1% of anomalies and 21% of normal data was falsely identified by our pipeline and re-evaluated as correctly labeled by the gastroenterologists.

The anomaly detection results in combination with the re-evaluation emphasizes that our cleaning pipeline is well-suited for mislabel identification. In the future, the sample size should be increased to re-evaluate a larger number of suspected samples.



(a) I: None - R: Angiectasia (b) I: None - R: Polyp (c) I: Blood - R: None (d) I: Polyp - R: None

Fig. 5: Representative VCE images, that were identified as mislabeled by our pipeline and re-annotated by clinical experts (I: Initial, R: Revised).

Furthermore, Figure 5 displays four representative VCE images identified by our proposed approach as mislabeled data, which were subsequently confirmed as incorrectly labeled by the scientific panel and re-annotated accordingly. The first two images were originally annotated as healthy mucosa, but have been corrected to angiectasia (Figure 5a) and polyp (Figure 5b). The second two frames were originally labeled as blood (Figure 5c) and polyp (Figure 5d), but were identified as normal samples by our pipeline and then relabeled respectively. This indicates the effectiveness of our pipeline in detecting present mislabeled data in medical datasets.

V. Conclusion

In this paper, we developed a machine learning-based pipeline to filter and correct medical datasets for potential mislabels. This was validated for the video capsule endoscopy, an essential real-world application to screen for diseases within the gastrointestinal tract. It was shown that anomaly detection for pathology identification in VCE images achieves superior results if applied to cleaned and filtered datasets, outperforming existing baselines that do not incorporate a prior data cleaning step by obtaining an accuracy of 93.83% and a F1-score of 71.58%. The effectiveness of this approach was further validated by cross-checking a small sample size by experienced physicians. Given the difficulties reported in previous publications regarding anomaly detection the proposed approach leads to improved results and constitutes a crucial contribution towards enabling on-device anomaly detection during video capsule endoscopies. In future work, all identified noisy samples may be re-annotated by gastroenterologists. Nevertheless, the dataset splits can already directly be used as labeled by our pipeline and are available here.

References

[1] Chiu, Y.C., Tsai, C.Y., Ruan, M.D., Shen, G.Y., Lee, T.T.: Mobilenet-ssdv2: An improved object detection model for embedded systems. In: 2020 International conference on system science and engineering (ICSSE). pp. 1–5. IEEE (2020)

[2] Christoph, G., Adrian, F., Tobias, H., Bernardo, P.P., Lübeck, K., Oliver, B.: Hardware accelerator and neural network co-optimization for ultra-low-power audio processing devices. In: 2022 25th Euromicro Conference on Digital System Design (DSD). pp. 365–369. IEEE (2022)

[3] Costamagna, G., Shah, S.K., Riccioni, M.E., Foschia, F., Mutignani, M., Perri, V., Vecchioli, A., Brizi, M.G., Picciocchi, A., Marano, P.: A prospective trial comparing small bowel radiographs and video capsule endoscopy for suspected small bowel disease. Gastroenterology 123(4), 999–1005 (2002)

[4] Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the em algorithm. Journal of the royal statistical society: series B (methodological) 39(1), 1–22 (1977)

[5] Feng, W., Quan, Y., Dauphin, G.: Label noise cleaning with an adaptive ensemble method based on noise detection metric. Sensors 20(23), 6718 (2020)

[6] Howard, A., Sandler, M., Chu, G., Chen, L.C., Chen, B., Tan, M., Wang, W., Zhu, Y., Pang, R., Vasudevan, V., et al.: Searching for mobilenetv3. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 1314–1324 (2019)

[7] Howard, A.G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., Adam, H.: Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861 (2017)

[8] Iddan, G., Meron, G., Glukhovsky, A., Swain, P.: Wireless capsule endoscopy. Nature 405(6785), 417–417 (2000)

[9] Jiang, G., Zhang, J., Bai, X., Wang, W., Meng, D.: Which is more effective in label noise cleaning, correction or filtering? In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 12866–12873 (2024)

[10] Kang, X., Duan, P., Xiang, X., Li, S., Benediktsson, J.A.: Detection and correction of mislabeled training samples for hyperspectral image classification. IEEE Transactions on Geoscience and Remote Sensing 56(10), 5673–5686 (2018)

[11] Le Floch, M., Wolf, F., McIntyre, L., Weinert, C., Palm, A., Volk, K., Herzog, P., Kirk, S.H., Steinhäuser, J.L., Stopp, C., et al.: Galar-a large multi-label video capsule endoscopy dataset. Scientific Data 12(1), 828 (2025)

[12] Lee, K.H., He, X., Zhang, L., Yang, L.: Cleannet: Transfer learning for scalable image classifier training with label noise. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5447–5456 (2018)

[13] Li, J., Socher, R., Hoi, S.C.: Dividemix: Learning with noisy labels as semi-supervised learning. arXiv preprint arXiv:2002.07394 (2020)

[14] Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE international conference on computer vision. pp. 2980–2988 (2017)
[15] Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
[16] McLachlan, G.J., Peel, D.: Finite mixture models. John Wiley & Sons (2000)
[17] Medtronic: PillCam™ SB3 System. https://www.medtronic.com/covidien/en-nz/products/capsule-endoscopy/pillcam-sb-3-system.html/ (2025), [Online; accessed 7-May-2025]
[18] Monteiro, S., de Castro, F.D., Carvalho, P.B., Moreira, M.J., Rosa, B., Cotter, J.: Pillcam® sb3 capsule: Does the increased frame rate eliminate the risk of missing lesions? World journal of gastroenterology 22(10), 3066 (2016)
[19] Northcutt, C.G., Wu, T., Chuang, I.L.: Learning with confident examples: Rank pruning for robust classification with noisy labels. arXiv preprint arXiv:1705.01936 (2017)
[20] Ostyakov, P., Logacheva, E., Suvorov, R., Aliev, V., Sterkin, G., Khomenko, O., Nikolenko, S.I.: Label denoising with large ensembles of heterogeneous neural networks. In: Proceedings of the European Conference on Computer Vision (ECCV) Workshops. pp. 0–0 (2018)
[21] Patrini, G., Rozza, A., Krishna Menon, A., Nock, R., Qu, L.: Making deep neural networks robust to label noise: A loss correction approach. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1944–1952 (2017)
[22] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in Python. Journal of Machine Learning Research 12, 2825–2830 (2011)
[23] Permuter, H., Francos, J., Jermyn, I.: A study of gaussian mixture models of color and texture features for image classification and segmentation. Pattern recognition 39(4), 695–706 (2006)
[24] Reed, S., Lee, H., Anguelov, D., Szegedy, C., Erhan, D., Rabinovich, A.: Training deep neural networks on noisy labels with bootstrapping. arXiv preprint arXiv:1412.6596 (2014)
[25] Sharma, K., Donmez, P., Luo, E., Liu, Y., Yalniz, I.Z.: Noiserank: Unsupervised label noise reduction with dependence models. In: European conference on computer vision. pp. 737–753. Springer (2020)
[26] Smedsrud, P.H., Thambawita, V., Hicks, S.A., Gjestang, H., Nedrejord, O.O., Næss, E., Borgli, H., Jha, D., Berstad, T.J.D., Eskeland, S.L., et al.: Kvasir-capsule, a video capsule endoscopy dataset. Scientific Data 8(1), 142 (2021)
[27] Swain, P., Iddan, G.J., Meron, G., Glukhovsky, A.: Wireless capsule endoscopy of the small bowel: development, testing, and first human trials. In: Biomonitoring and Endoscopy Technologies. vol. 4158, pp. 19–23. SPIE (2001)
[28] Tanaka, D., Ikami, D., Yamasaki, T., Aizawa, K.: Joint optimization framework for learning with noisy labels. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2018)
[29] Thomson, A., Keelan, M., Thiesen, A., Clandinin, M., Ropeleski, M., Wild, G.: Small bowel review: diseases of the small intestine. Digestive diseases and sciences 46, 2555–2566 (2001)
[30] Wang, Y., Ma, X., Chen, Z., Luo, Y., Yi, J., Bailey, J.: Symmetric cross entropy for robust learning with noisy labels. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 322–330 (2019)
[31] Werner, J., Bause, O., Oexle, J., Floch, M.L., Brinkmann, F., Hampe, J., Bringmann, O.: Seeing more with less: Video capsule endoscopy with multi-task learning. arXiv preprint arXiv:2507.23479 (2025)
[32] Werner, J., Gerum, C., Nick, J., Floch, M.L., Brinkmann, F., Hampe, J., Bringmann, O.: Enhanced anomaly detection for capsule endoscopy using ensemble learning strategies. arXiv preprint arXiv:2504.06039 (2025)
[33] Werner, J., Gerum, C., Reiber, M., Nick, J., Bringmann, O.: Precise localization within the gi tract by combining classification of cnns and time-series analysis of hmms. In: International Workshop on Machine Learning in Medical Imaging. pp. 174–183. Springer (2023)
[34] Xiao, T., Xia, T., Yang, Y., Huang, C., Wang, X.: Learning from massive noisy labeled data for image classification. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2691–2699 (2015)
[35] Zhang, Z., Sabuncu, M.: Generalized cross entropy loss for training deep neural networks with noisy labels. Advances in neural information processing systems 31 (2018)

Smart Video Capsule Endoscopy: Raw Image-Based Localization for Enhanced GI Tract Investigation*

Oliver Bause[0009−0003−5388−2959]†, Julia Werner[0009−0006−0279−1776]†, Paul Palomero Bernardo[0000−0002−6642−3976], and Oliver Bringmann[0000−0002−1615−507X]

Chair of Embedded Systems, Eberhard Karls University of Tübingen, 72076 Tübingen, Germany
oliver.bause@uni-tuebingen.de
julia-helga.werner@uni-tuebingen.de
https://www.embedded.uni-tuebingen.de/en/home/

Abstract. For many real-world applications involving low-power sensor edge devices deep neural networks used for image classification might not be suitable. This is due to their typically large model size and requirement of operations often exceeding the capabilities of such resource limited devices. Furthermore, camera sensors usually capture images with a Bayer color filter applied, which are subsequently converted to RGB images that are commonly used for neural network training. However, on resource-constrained devices, such conversions demands their share of energy and optimally should be skipped if possible. This work addresses the need for hardware-suitable AI targeting sensor edge devices by means of the Video Capsule Endoscopy, an important medical procedure for the investigation of the small intestine, which is strongly limited by its battery lifetime. Accurate organ classification is performed with a final accuracy of 93.06% evaluated directly on Bayer images involving a CNN with only 63,000 parameters and time-series analysis in the form of Viterbi decoding. Finally, the process of capturing images with a camera and raw image processing is demonstrated with a customized PULPissimo System-on-Chip with a RISC-V core and an ultra-low power hardware accelerator providing an energy-efficient AI-based image classification approach requiring just 5.31 µJ per image. As a result, it is possible to save an average of 89.9% of energy before entering the small intestine compared to classic video capsules.

Keywords: Wireless Capsule Endoscopy · Bayer Image Classification · Smart Edge Devices.

† Equal contribution: These authors contributed equally to this work.

* This work has been partly funded by the German Federal Ministry of Research, Technology and Space (BMFTR) in the projects MEDGE (16ME0530) and Scale4Edge (16ME012).

1 Introduction

Vision-based deep neural networks are commonly trained with standard RGB images. While cameras typically capture images in Bayer pattern format, they are normally converted to RGB images afterwards for further utilization. In the field of machine learning, this circumstance is often not addressed since, for many applications, conducting such a conversion can be easily performed. However, for a range of real-world applications relying on small, low power edge devices this conversion is costly and should be avoided to save the limited power if possible.

Video Capsule Endoscopy (VCE) is an example for a medical application that requires an edge device with limited power availability [13,23]. This medical procedure was first introduced in the early 2000s and comprises the swallowing of a small pill-sized capsule with LEDs and a camera that can capture images of the digestive tract while the capsule moves from the esophagus through the stomach and small intestine to the colon. Subsequently, the images are transmitted to an external on-body receiver for further inspection by qualified physicians. This examination of the gastrointestinal (GI) tract is crucial for the investigation of the otherwise largely inaccessible small intestine which can exhibit gastroenterological pathologies such as ulcers, inflammation, polyps or cancer. Due to the limited volume of the capsule, the overall battery lifetime is severely restricted. For example, for the PillCam™ SB3 from Medtronic an operation time of 8 to 12 h is reported with 2 to 6 frames per second (fps) [18,19]. However, the time of passing through the whole GI tract varies from patient to patient and can easily exceed 12 h in certain cases. Given the human anatomy of the GI tract, this can result in an incomplete screening which is especially problematic if anomalies or diseases are located in the uncovered region of interest.

Our Contribution: To prolong the battery lifetime of small edge devices, that incorporate artificial intelligence (AI) for image classification, this work aims to reduce the energy consumed per inference by using raw Bayer instead of the commonly used RGB images. This enables a precise localization of the capsule within the GI tract to determine the moment when the region of interest, the small intestine, is entered. Images obtained from the preceding organs are discarded and the transmission only starts after this time point. Furthermore, the frame rate before entering the small intestine can be reduced which also saves power. Moreover, first experiments on training a very light-weight convolutional neural network (CNN) directly on images in Bayer pattern format are conducted which saves an additional conversion step on hardware and also provides a baseline for future experiments. The whole process is demonstrated using a customized PULPissimo System-on-Chip (SoC) with a RISC-V core and an ultra-low power hardware accelerator. This process consits of capturing VCE images with a miniature camera, analyzing them with a small CNN, and parsing the labels to a microcontroller unit (MCU) for post-processing with Viterbi decoding. To conclude, we provide an energy-efficient AI classification pipeline for vision applications and demonstrate this with the example of VCE.

2 Related Work

2.1 CNN Training on Bayer images

Mobile devices are generally equipped with camera sensors that capture images with a Bayer color filter applied. To improve the quality of these images, it is necessary to demosaic and denoise them. Research is ongoing to enhance the post-processing by leveraging neural networks [14, 15]. However, the integration of such networks into ultra-low power devices is not a viable solution, as these are characterized by high computational demands. Additionally, CNNs employed for vision tasks are typically trained on three-channel RGB images rather than raw, mosaicked Bayer images. [8] investigated the potential of training SqueezeNet with Bayer images to classify hand postures. The accuracy of the method was determined to be 98.28%, with a co-sited demosaicing technique employed. However, when using raw Bayer images, the accomplished accuracy was less than 25%.

2.2 Hardware-Suitable CNNs for Video Capsule Endoscopies

Precise organ classification based on VCE images has been substantially progressed due to the release of a publicly available VCE dataset, the Rhode Island Gastroenterology dataset [9], with annotations for the four main organs of the GI tract, which are traversed by a video capsule. [9] provides a baseline with an Inception ResNetV2, achieving an accuracy of 97.1 %. This problem was further addressed by [24] and [1], yielding accuracies of 96.95 % and 69.84 %, respectively. While [24] aimed to lower the model complexity and constrain the total number of parameters to 1M, the proposed approaches still demand too many resources from potential hardware architectures. [20] provided first results on a programmable edge AI accelerator which can efficiently execute a simple convolutional network while demonstrating proficient results on the VCE task. Importantly, all those findings are based on CNN evaluations using RGB images. [21] reported first results on real-time VCE image processing with onboard neural networks. However, the neural network employed, still consisted of $\approx$ 3 M parameters and their prototype did not exceed an operation time of 1 h, limiting the applicability due to a high power consumption.

Based on this previous research, we aim to perform neural network training directly on images in Bayer pattern format to omit the conversion step to the RGB format combine the CNN evaluations with post-processing with an HMM and Viterbi decoding as shown by [24] with additional quantization and finally demonstrate this with a hardware-in-the-loop (HIL) testing [4].

3 Methodology

The proposed method is illustrated in Figure 1 involving the training of a quantized CNN which receives VCE images in raw Bayer pattern format as an input

and parses its evaluations to a quantized HMM. Since the camera forwards images directly in Bayer pattern format and conversion to RGB images requires it's share of energy and memory, the dataset at hand is converted to Bayer pattern frames to simulate the conditions of this medical application realistically. Additionally, the capsule's composition is simulated with an asm NanEyeC miniature camera that sends images in Bayer pattern format to a hardware accelerator that can efficiently execute the CNN and features an MCU for subsequent processing of quantized Viterbi decoding. During the application, this decision pipeline should help to decide, when to start transmission of VCE frames to an on-body device and adjust the frame rate to lower the overall power consumption. Finally, only when outside of the capsule, the images are converted to RGB format for evaluations by medical experts.

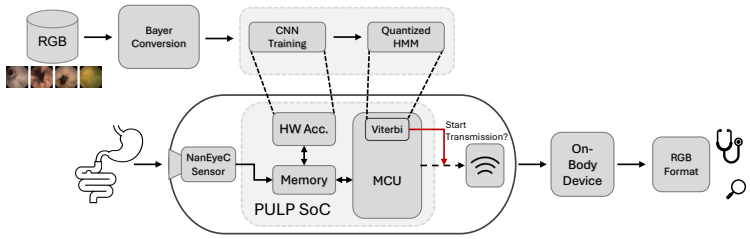


Fig. 1: Proposed method of processing raw input VCE images with a quantized low-complex CNN and quantized Viterbi decoding on a VCE demonstrator involving the NaneyeC camera, a hardware accelerator and an MCU.

3.1 Bayer Filter Conversion

Image sensors, especially those with a miniature form factor, capture images with a color filter array (CFA) applied [2,5]. The Bayer filter is the most common one and uses a 2×2 pattern with two green, one red, and one blue pixel. The resulting filters are called BGGR, RGBG, GRGB, and RGGB. Thus, Bayer images are space-saving as they only have one color channel per pixel. To obtain an RGB image from a raw Bayer image, the missing color channels of each pixel need to be calculated by interpolating the corresponding color channels of the neighboring pixels. Further demosaicing can then be performed to increase image quality.

Images within datasets are normally stored in an RGB image format like PNG or JPEG. Nevertheless, these were often captured using a sensor with a CFA. Consequently, it is feasible to convert them back into the original format with minimal loss of data by extracting the original color filter again.

3.2 Rhode Island Gastroenterology Dataset

Training neural networks for performing localization within the GI tract based on VCE images requires an extensive database of VCE studies. The Rhode Island gastroenterology dataset [9] is one of the largest publicly available VCE datasets and consists of $5,247,588$ labeled images from 424 VCE studies in total. They feature annotations for the four organs: esophagus, stomach, small intestine and colon. In this work, this dataset along with its official splits of training, validation and test set was adopted for all further experiments.

3.3 CNN Architectures and Training Specifications

Neural networks from the Mobilenet [12] family have been specifically designed to target applications for mobile devices and are thus less complex than other common neural networks (e.g. large ResNets or vision transformer). Furthermore, these model architectures have been used for the problem of classifying low-resolution images from VCEs recently [24]. Following those findings, we first employed a MobileNetV3-Small architecture in 32-bit floating point representation which was pretrained on Imagenet to explore the classification performance of this model with standard RGB images as an input compared to raw Bayer pattern images using the same training conditions.

However, considering that this model still consists of 1M parameters in total, lowering the model complexity is still desirable. Thus, an even less complex, residual network that has been recently introduced by [20], was adapted and tested with RGB images as well as inputs in raw Bayer pattern format. This model consists of less than $100,000$ parameters and has been successfully deployed on an AI hardware accelerator with high energy efficiency. All models were trained for 10 epochs with a learning rate of 0.001 and a weight decay of 10^{-5} using the AdamW optimizer in PyTorch [3].

3.4 Hidden Markov Model and Viterbi Decoding

As shown in [24], the path of such video capsule can be modeled by a Hidden Markov Model (HMM) in which each organ is represented as a hidden state and the transition probabilities mark the transitioning into a new organ. The emission probabilities model the probability of observing a certain organ while actually being in a specific organ at a moment t. Subsequent Viterbi decoding [11] returns the most likely path of organs based on the CNN evaluations. We adapted this approach with few modifications to ensure hardware suitability. As conducted by [24], we compute the Viterbi algorithm using solely the log likelihoods, which only require additions instead of multiplications as operations. This prevents numerical instability in case of very small likelihoods. However, instead of floating-point operations, we employ strict fixed point representation and quantize not only the weights, bias and features of the CNN but, additionally, generate a quantized Viterbi algorithm in Python as well as C, that can be directly transferred to the following hardware architecture.

3.5 System Architecture

The system architecture of our VCE demonstrator relies on a customized PULPissimo SoC which features a RISC-V core [7,22]. The chip can be clocked between 20 to 400 MHz and features 384 KiB of SRAM. A stringent clock-gating approach which involves regulating the core, the accelerator, and the memories enables the utilization for energy-constrained applications.

The SoC has been extended with an integrated hardware controller to efficiently communicate with the asm NanEyeC miniature camera sensor [2]. The camera captures images with a resolution of 320×320 pixels at up to 58fps with a RGGB CFA applied and a size of just $1\,\mathrm{mm}^2$. The controller handles the configuration and communication with the camera autonomously to off-load the RISC-V core and to save energy. Compared to the camera's generic SPI protocol, the controller also formats and stores the received image data as packed `uint8` in the SoC's memory independently.

Our custom PULPissimo is equipped with the UltraTrail AI accelerator [6] for real-time, ultra-low power inference of temporal convolutional networks. The accelerator enables the SoC to perform on-the-edge AI image classification with 64 Multiply-and-Accumulate (MAC) units. UltraTrail directly accesses the SoC's memory banks via the TCDM interconnect to retrieve the features and weights and to store the results of the current layer. Thus, it only requires a dedicated 136 KiB SRAM for interim results.

The PULPissimo SoC is synthesized in `22FDX+` technology and has a chip area of only 2 mm by 1.8 mm. Simulations are performed to obtain energy consumption estimations of a taped-out design that can be integrated into a VCE prototype. However, to verify the functionality of the proposed architecture, the chip is programmed onto the Digilent Nexys A7-100T FPGA [10]. This allows us to simulate the whole system within a controlled environment. The FPGA is connected to the HIL setup presented in [4]. The HIL offers a digital twin of the NanEyeC camera sensor and is connected to a database containing the Rhode Island [9] test set. Thus, it is possible to simulate the traversal of the VCE prototype through the studies' GI tracts. The accelerator's output is verified against our pre-trained model and the original labels of the data set. Various capsule settings can be tested in real-time to find the optimal configuration in regard to energy consumption, frame rate and latency. Additionally, functional correctness can be assured before performing time- and cost-intensive animal testing, reducing the overall development cycle.

The firmware executed on the PULPissimo SoC has the task of utilizing the modules and sensors of the capsule in order to achieve the best possible screening of the area of the GI tract to be analyzed. The traversal through the entire GI tract can take up more than 12 h. Therefore, the available energy must be utilized in a carefully targeted manner. The firmware captures images with a dynamic frame rate depending on the current system state. The images are then analyzed by the hardware accelerator and a memory-efficient quantized implementation of the Viterbi decoding to determine the position of the capsule. As long as it is still

not in the small intestine, the frame rate can be reduced and the transmission of the images to the on-body receiver can be deactivated to further save energy.

4 Results

4.1 Classification Performance: RGB - Bayer

All baseline studies for this dataset were conducted using RGB images. However, as previously mentioned, converting the initial Bayer pattern format to the standard RGB image, demands its share of energy. Thus, the question arises, how well a CNN can process and classify images in Bayer pattern format compared to RGB images. As a first step, we compared the light-weight MobileNet trained on either RGB or Bayer pattern images in comparison to the baseline results from the literature (see Table 1).

Table 1: Comparison of the light-weight MobileNet trained on either RGB or Bayer pattern images (RGGB) in comparison to literature results.

	Input	**Accuracy**	**Recall**	**Precision**	**F1-Score**	**Params**
Inception ResNetV2 [9]	RGB	97.1	97.07	97.31	97.13	56 M
Swin Transformer [1]	RGB	69.84	69.91	69.72	69.85	195 M
MobileNetV3 [24]	RGB	96.95	–	–	–	1 M
MobileNetV3	RGB	97.14	95.74	87.25	90.89	1 M
MobileNetV3	RGGB	96.20	94.30	83.92	88.12	1 M

All previous experiments on this dataset were conducted with RGB images as an input, which functions as the standard in the field of computer vision. Based on [24], we adapt the usage of the low-complex MobileNet architecture for this image classification problem achieving a similar accuracy of 97 % (compared to 96.95%) with a total of 1 M parameters. Compared to the baseline [9] with an accuracy of 97.1 % and an F1-Score of 97.13 %, the utilized model performs slightly worse with an accuracy of 97.14 % and a F1-Score of 90.89 %, however, requiring only 1 M instead of 56 M parameters. Furthermore, the employed model needs drastically less parameters (1 M vs 195 M) compared to [1], while outperforming the Swin transformer notably (F1-Score of 90.89 % vs. 69.85 %). MobileNet already functions as a hardware-aware network which was specifically designed for small mobile devices. For ultra-low power applications like VCE, this is still too complex and thus 1 M parameters would require a large area of a chosen hardware architecture. We can further observe that the evaluation on RGGB images only marginally reduces classification performance. Compared to RGB inputs, the accuracy is only reduced from 97.14% to 96.20% while precision

and F1-Score decrease more significantly. However, these findings indicate that employing a light-weight model directly on RGGB images as received from a camera still allows accurate image classification.

Thus, based on these results, we explored if an even less complex model with only 62,976 parameters can also capture the class differences if RGGB images are received as an input in comparison to the standard RGB images. Figure 2 shows a comparison of the classification performance of the presented hardware-suitable CNN with RGGB images as input compared to standard RGB images over different word widths.

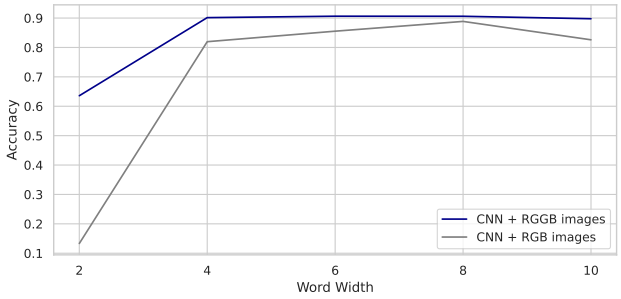


Fig. 2: Classification performance of the hardware-suitable CNN over different word widths, validated either with raw Bayer pattern (RGGB) images or RGB images as input.

This demonstrates that this model is capable of classifying VCE images with a high accuracy even if directly executed on images in Bayer pattern format, without requiring a preceding conversion to RGB format. Surprisingly, directly training on RGGB images seemingly performs better than using standard RGB images with this network.

4.2 Postprocessing with Viterbi Decoding

In the following, we aim to enhance the classification ability of this network even further by combining the CNN with an HMM and subsequent Viterbi decoding (see Table 2).

Combining the localization net with an HMM enhances the classification performance to an accuracy of 93.06 % and a sensitivity of 89.55 %. While this model performs inferior compared to the baseline results, it still provides remarkable classification capabilities considering the low complexity. The presented approach only requires 2,874,880 MAC operations and a total of 62,976 parameters, instead of 1 M. Since during Viterbi decoding a number of observations

Table 2: Classification performance of the hardware-suitable CNN (word width=8 bit) in combination with HMM and Viterbi Decoding (window size = 50).

	Accuracy	**Recall**	**Precision**	**F1-Score**	**Params**
Localization Net	90.61	83.98	71.49	75.81	62,976
Localization Net & HMM	93.06	89.55	80.65	84.08	62,976

need to be received to estimate the most likely point in time for observing a specific state (with the VCE as an example: the small intestine), employing this method can be accompanied by a certain delay. Thus, for all patients within the test set the delay, meaning the difference between the actual first occurrence of the small intestine and the detection by the Viterbi decoding, is depicted in Figure 3.

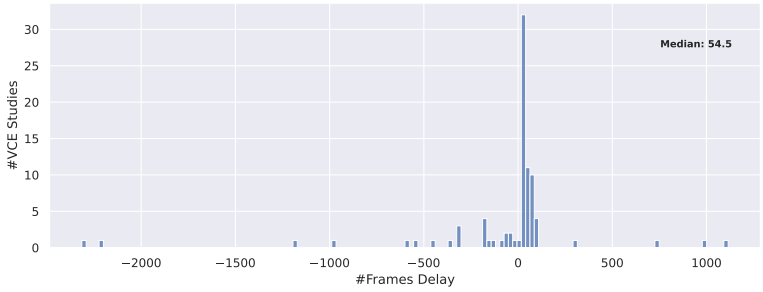
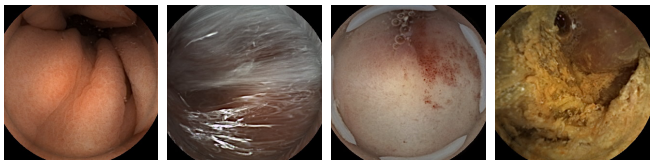


Fig. 3: Distribution of delays across all VCE studies.

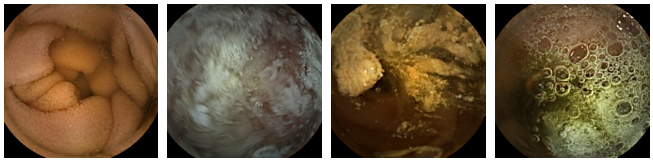
With a median delay of 54.5, the plot visualizes that the HMM predicts the entering of the small intestine accurately for the majority of patients. Only a few outlier studies resulted in extreme delays (either too early or too late detection), which are investigated closer in the following.

In Figure 4, for the stomach as well as the small intestine, example VCE frames are shown with standard images for each organ in 4a and 4e as a reference and, in comparison, examples VCE images from the stomach 4b-4d and from the small intestine 4f-4h are displayed for outlier studies with the largest obtained delays.

The presented VCE example images display various obstacles such as bubbles and digestion remains. The VCE study with the ID 23 had a delay of 2180, study



(a) Standard ST (b) ID18-ST (c) ID23-ST (d) ID193-ST



(e) Standard SI (f) ID18-SI (g) ID193-SI (h) ID221-SI

Fig. 4: Examples of VCE images from the Rhode Island [9] test set with frames originating from the stomach (ST) on top and frames from the small intestine (SI) on the bottom. (a) and (e) display a common image for each organ as present in the dataset. In comparison, (b)-(d) and (f)-(h) are from the studies resulting in the lowest accuracies or delays in the presented work.

18 a delay of 1915. For studies 221 and 193, the small intestine was detected 4680 and 4496 frames too early respectively. The visual differences as shown in Figure 4 seemingly challenge the image detection process.

4.3 Demonstrator Performance

As the battery capacity is constrained by the size of the capsule, the whole system cannot be active all the time transmitting 58 frames per second to the on-body receiver. As shown in Table 3, image capturing is the main energy consumer within the image processing pipeline. Additionally, the transmission of images to the on-body receiver also requires its share of energy.

The receiver typically draws 5 mW and achieves 2 MiBps [17, 21]. An image has a size of 102.4 KiB, thus requires 50 ms to be transmitted to the on-body receiver. The resulting energy consumed per frame is ∼250 µJ. Thus, reducing the amount of worthless images captured and transmitted increases the battery life of the VC. By incorporating UltraTrail and a quantized implementation of the HMM, the captured images can be analyzed to determine the current location within the GI tract. The energy consumed to process the images is at 0.31 µJ negligible compared to the image sensor, the LEDs, and the transmitter module. Thus, the firmware captures images with a defined frame rate and analyzes them with the HW accelerator in combination with Viterbi decoding. The capsule's arrival in the small intestine, as indicated by the HMM, initiates the actual

Table 3: Power (mW) and energy (µJ) consumption by each module per frame processed with comparison to the AI capsule of [21]. Metrics that are not mentioned in their paper are marked as undefined (U). Note that no energy consumption is listed for the idle state of our demonstrator, as the time in idle depends on the current frame rate, thus it is not constant.

Task	Image Capture			DNN	HMM	Idle	
Module	**Image Sensor**	**LEDs**	**MCU**	**MCU with acc.**	**MCU**	**MCU**	**Image Sensor**
Power Consumption of [21] [mW]	~ 60	U	U	~ 50	-	N/A	0.1
Power Consumption [mW]	8.51	14.78	7.23	16.63	9.94	5.85	3.0
Inference Time [ms]	12.79	12.79	12.79	0.31	0.02	-	-
Energy Consumption [µJ]	108.93	189.15	92.56	5.14	0.17	-	-

screening process thereby increasing the frame rate and triggering the subsequent transmission of image data.

Simulation of GI Tract Traversals: To verify the functionality of our demonstrator and evaluate its performance, multiple studies from the Rhode Island test set are simulated by the HIL setup [4]. The studies were captured by using the PillCamTM SB3 capsule, which utilizes a similar image sensor with a resolution of 320×320 pixels and a dynamic frame rate of 2 to 6 fps. In order to ensure the comparability of different simulations, it is assumed that the frame rate is fixed at 2 fps during the study's original recording. The objective of on-edge location detection is to identify the point of entering the area of interest, the small intestine, with minimal delay while simultaneously spending only as much energy as absolutely necessary. A delay that is less than zero indicates that the transmission is initiated prematurely. This results in the consumption of energy that could be allocated for more crucial images. Conversely, a high delay suggests that the transmission is initiated after the optimal time, resulting in the skipping of segments at the beginning of the small intestine. As the standard endoscopic examination of the gastroscopy is capable of screening up to 30 cm of the small intestine [16], a slight positive delay is permissible. The baseline of the experiment is an ordinary video capsule, that captures 2 fps and transmits them without further processing to the on-body receiver. Thus, the baseline has a deeply negative delay as it sends all frames, also those of the esophagus and stomach.

To illustrate the capabilities of the location detection, study 73 of the Rhode Island test set is used as exemplary screening. As shown in Figure 5, the baseline capsule spends 1140 mJ of energy prior to attaining the area of interest, which is effectively dissipated. Using the same frame rate and analyzing the images on-

edge, however, only requires 709 to 816 mJ even though a small HMM window size leads to a premature misdetection of the small intestine, which starts the image transmission too early.

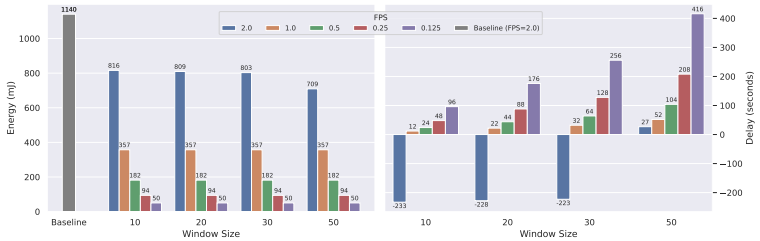


Fig. 5: Energy consumption (left) and delay in frames (right) of different HMM window size and fps combinations for the simulation of study 73.

This misdetection is initiated by a series of successive images of compromised quality (Figure 6, DNN Predictions), which collectively challenge the capabilities of the CNN. Therefore, a decrease in the frame rate has two primary benefits. It reduces energy consumption and it also decreases the likelihood of consecutive frames containing dirt.

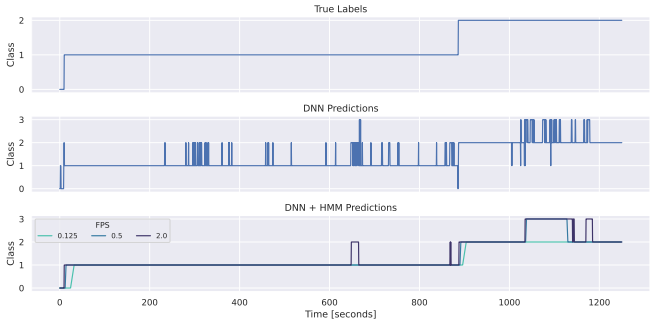
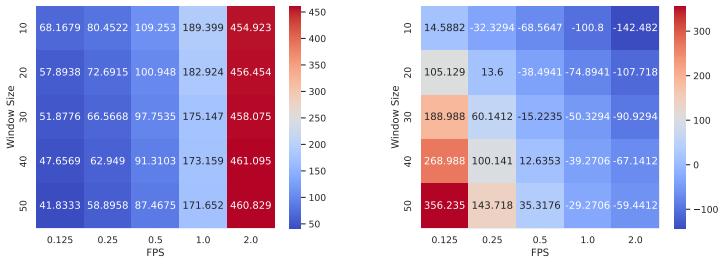


Fig. 6: Original labels, DNN predictions, and HMM prediction of different fps with a window size of 10 for study 73 (with the classes: 0 - Esophagus, 1 - Stomach, 2 - Small intestine, 3 - Colon).

It should be noted that this process concomitantly engenders an increase in the delay between the transition into the small intestine and the HMM detecting the transition as more time elapses before the window is filled with a path pointing towards a small intestine prediction. Thus, a trade-off between frame rate and HMM window size needs to be found to achieve a low energy consumption while keeping the delay close to zero. Figure 7 displays the results of a hyperparameter grid search that was performed on all Rhode Island test studies. The baseline achieved an average energy consumption across all studies of 719.934 mJ.



(a) Average Energy Consumption [mJ] (avg. of baseline: 719.934 mJ)

(b) Average Delay [s]

Fig. 7: Grid search of the HMM hyperparameters window size and frame rate to find the optimal combination in regard of average energy consumption (a) and average delay (b) across all test studies of the Rhode Island data set.

As expected, the average energy consumed increases with a higher frame rate. For example, an increase in frame rate from 1 to 2 fps multiplies the energy consumed by $\sim 240\%$ since a higher frame rate can result in an early detection of the small intestine, because of multiple consecutive images of a same scene that challenges the CNN, and thus starts the transmission prematurely. On the other hand, a very low frame rate combined with a large window size is energy-saving but it inserts a massive average delay of more than 350 s in which the capsule may travel further through the small intestine than the 30 cm that can be screened using the standard endoscopic examination. As a result, crucial segments of the small intestine may not be screened at all. A frame rate of 0.25 fps with a window size of 20 seems to be a good trade-off between energy consumption and delay and saves $\sim 89.9\%$ energy compared to the baseline capsule. If more energy is available, 0.5 fps with a window size of 30 or even 40 could also be possible to make post-processing more stable against some mispredictions of the CNN.

5 Conclusion

In the presented work, extensive simulation from capturing images with a NanEyeC miniature camera sensor to final evaluation with a hardware suitable approach composing of a light-weight CNN and subsequent quantized Viterbi decoding is shown. By training the model directly on raw Bayer images, an additional costly conversion step implementation on the system's architecture is avoided. While requiring only 62,976 parameters for the CNN, an accuracy of 93.06 % is achieved. The power and energy consumption for each module is reported for this setup, yielding a total energy consumption of only 5.14 µJ for the DNN and 0.17 µJ for the Viterbi decoding per frame analyzed. A grid search demonstrated that it is advisable to decrease the frame rate to 0.25 fps with a HMM window size of 20 to achieve a 89.9% lower power consumption while still detecting the transition into the small intestine timely. It is demonstrated that the presented setup is capable of real-time on-edge location detection.

References

1. Abian, A.I., Raiaan, M.A.K., Jonkman, M., Islam, S.M.S., Azam, S.: Atrous spatial pyramid pooling with swin transformer model for classification of gastrointestinal tract diseases from videos with enhanced explainability. Engineering Applications of Artificial Intelligence **150**, 110656 (2025)
2. ams-OSRAM AG: NanEyeC Miniture Camera Module (Oct 2024), v5-00, https://look.ams-osram.com/m/19863b4335e67f1b/original/NanEyeC-Miniature-Camera-Module.pdf
3. Ansel, J., et al.: PyTorch 2: Faster Machine Learning Through Dynamic Python Bytecode Transformation and Graph Compilation. In: 29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2 (ASPLOS '24). ACM (Apr 2024). https://doi.org/10.1145/3620665.3640366, https://pytorch.org/assets/pytorch2-2.pdf
4. Bause, O., Werner, J., Bringmann, O.: Systematic hardware integration testing for smart video-based medical device prototypes. In: 2025 47th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). pp. 1–6. IEEE (2025)
5. Bayer, B.E.: Color imaging array. https://patents.google.com/patent/US3971065/ (7 1976)
6. Bernardo, P.P., Gerum, C., Frischknecht, A., Lübeck, K., Bringmann, O.: Ultra-trail: A configurable ultralow-power tc-resnet ai accelerator for efficient keyword spotting. IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems **39**(11), 4240–4251 (2020)
7. Bernardo, P.P., Schmid, P., Bringmann, O., Iftekhar, M., Sadiye, B., Mueller, W., Koch, A., Jentzsch, E., Sauer, A., Feldner, I., et al.: A scalable risc-v hardware platform for intelligent sensor processing. In: 2024 Design, Automation & Test in Europe Conference & Exhibition (DATE). pp. 1–5. IEEE (2024)
8. Chandra, M., Lall, B.: A novel method for cnn training using existing color datasets for classifying hand postures in bayer images. SN Computer Science **2**(2), 60 (2021)

9. Charoen, A., Guo, A., Fangsaard, P., Taweechainaruemitr, S., Wiwatwattana, N., Charoenpong, T., Rich, H.G.: Rhode island gastroenterology video capsule endoscopy data set. Scientific Data **9**(1), 602 (2022)
10. Digilent Inc.: Nexys A7 FPGA Board Reference Manual (Jul 2019), https://digilent.com/reference/programmable-logic/nexys-a7/start
11. Forney, G.D.: The viterbi algorithm. Proceedings of the IEEE **61**(3), 268–278 (1973)
12. Howard, A., Sandler, M., Chu, G., Chen, L.C., Chen, B., Tan, M., Wang, W., Zhu, Y., Pang, R., Vasudevan, V., et al.: Searching for mobilenetv3. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 1314–1324 (2019)
13. Iddan, G., Meron, G., Glukhovsky, A., Swain, P.: Wireless capsule endoscopy. Nature **405**(6785), 417–417 (2000)
14. Khadidos, A.O., Khadidos, A.O., Khan, F.Q., Tsaramirsis, G., Ahmad, A.: Bayer image demosaicking and denoising based on specialized networks using deep learning. Multimedia Systems **27**(4), 807–819 (2021)
15. Kumar, S.P., Peter, K.J., Kingsly, C.S.: De-noising and demosaicking of bayer image using deep convolutional attention residual learning. Multimedia Tools and Applications **82**(13), 20323–20342 (2023)
16. Lewis, B.S.: Endoscopy, pp. 29–37. Springer International Publishing, Cham (2015). https://doi.org/10.1007/978-3-319-14415-3_3, https://doi.org/10.1007/978-3-319-14415-3_3
17. Liu, G., Yan, G., Zhu, B., Lu, L.: Design of a video capsule endoscopy system with low-power asic for monitoring gastrointestinal tract. Medical & biological engineering & computing **54**, 1779–1791 (2016)
18. Medtronic: PillCam™ SB3 System. https://www.medtronic.com/covidien/en-nz/products/capsule-endoscopy/pillcam-sb-3-system.html/ (2025), [Online; accessed 7-May-2025]
19. Monteiro, S., de Castro, F.D., Carvalho, P.B., Moreira, M.J., Rosa, B., Cotter, J.: Pillcam® sb3 capsule: Does the increased frame rate eliminate the risk of missing lesions? World journal of gastroenterology **22**(10), 3066 (2016)
20. Palomero Bernardo, P., Schmid, P., Gerum, C., Bringmann, O.: Compiler-aware ai hardware design for edge devices. In: Proceedings of the 8th International Workshop on Edge Systems, Analytics and Networking. pp. 31–36 (2025)
21. Sahafi, A., Wang, Y., Rasmussen, C., Bollen, P., Baatrup, G., Blanes-Vidal, V., Herp, J., Nadimi, E.: Edge artificial intelligence wireless video capsule endoscopy. Scientific reports **12**(1), 13723 (2022)
22. Schiavone, P.D., Rossi, D., Pullini, A., Di Mauro, A., Conti, F., Benini, L.: Quentin: an ultra-low-power pulpissimo soc in 22nm fdx. In: 2018 IEEE SOI-3D-Subthreshold Microelectronics Technology Unified Conference (S3S). pp. 1–3 (2018). https://doi.org/10.1109/S3S.2018.8640145
23. Swain, P., Iddan, G.J., Meron, G., Glukhovsky, A.: Wireless capsule endoscopy of the small bowel: development, testing, and first human trials. In: Biomonitoring and Endoscopy Technologies. vol. 4158, pp. 19–23. SPIE (2001)
24. Werner, J., Gerum, C., Reiber, M., Nick, J., Bringmann, O.: Precise localization within the gi tract by combining classification of cnns and time-series analysis of hmms. In: International Workshop on Machine Learning in Medical Imaging. pp. 174–183. Springer (2023)

List of Abbreviations

1D 1-Dimensional

2D 2-Dimensional

AI Artificial Intelligence

AUC Area Under the Curve

CCD Charge-Coupled Devices

CFA Color Filter Array

CHB-MIT Children's Hospital Boston - Massachusetts Institute of Technology

CMOS Complementary Metal Oxide Semiconductor

CNN Convolutional Neural Network

DBS Deep Brain Stimulation

DNA Deoxyribonucleic Acid

DWA Dynamic Weight Average

EDF European Data Format

EEG Electroencephalography

EWMA Exponentially Weighted Moving Average

FDA Food and Drug Administration

FP False Positives

GI Gastrointestinal

GMM Gaussian Mixture Model

GPU Graphics Processing Units

h hour

HMM Hidden Markov Model

IBD Inflammatory Bowel Disease

IFM Input Feature Map

LED Light-Emitting Diodes

LSTMs Long Short-Term Memory

M Million

MAC Multiply-Accumulate

MCU Microcontroller Unit

MIMD Multiple Instruction Streams Multiple Data Streams

MISD Multiple Instruction Streams Single Data Stream

ML Machine Learning

MSE Mean Squarred Error

NAS Neural Architecture Search

PCA Principal Component Analysis

pp percentage points

PSSIM Part-Based Structural Similarity Index Measure

ReLU Rectified Linear Unit

ResNet Residual Networks

RGB Red Green Blue

RGGB Red Green Green Blue

RI Rhode Island

RISC Reduced Instruction Set Computer

RNN Recurrent Neural Networks

RNS Responsive Stimulation

SAD Semi-Supervised Anomaly Detection

SB Small Bowel

SIMD Single Instruction Multiple Data Streams

SISD Single Instruction Stream Single Data Stream

SMA Simple Moving Average

SoC System-on-Chip

SSIM Structural Similarity Index Measure

SVM Support Vector Machines

t-SNE t-Distributed Stochastic Neighbor Embedding

TCN Temporal Convolutional Neural Networks

TP True Positives

TUH Temple University Hospital

VCE Video Capsule Endoscopy

VNS Vagus Nerve Stimulation

List of Figures

List of Tables