Uncertainty-aware learning from sparse, unlabelled, and out-of-distribution time series

Sotirios Vavaroutas

St Edmund's College

December 2025

This thesis is submitted for the degree of Doctor of Philosophy

Declaration

This thesis is the result of my own work and includes nothing which is the outcome of work done in collaboration or content generated by artificial intelligence, except as declared in the preface and specified in the text. It is not substantially the same as any work that has already been submitted, or is being concurrently submitted, for any degree, diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the preface and specified in the text. It does not exceed the prescribed word limit for the relevant Degree Committee.

Abstract

Nowadays, machine learning is increasingly popular in the analysis of healthcare time series, as it can support improved diagnostics, personalised monitoring, and effective performance tracking. The ever-increasing availability of datasets from wearable sensors, mobile devices, and continuous monitoring technologies has created new opportunities for real-world applications, as they can offer a rich and continuous picture of an individual's health and fitness in everyday environments as well as in clinical or supervised healthcare settings.

Despite the potential that machine learning models offer for healthcare time series, they still face notable challenges. Sensor-based datasets are frequently sparse (with missing values) if acquired outside controlled environments, while a substantial proportion remain unlabelled or only partially labelled. Further, models often exhibit limited generalisability under out-of-distribution conditions, which arise due to heterogeneity across hospitals, sensors, and patient populations. Such factors limit the use of real-world time series in deep learning applications, and are consequential in healthcare as unreliable models may not only yield poor performance but also produce overconfident metrics.

In this context, this thesis moves beyond the narrow settings of conventional models through uncertainty-aware learning. Effectively incorporating uncertainty enhances training by addressing key challenges, as it informs the handling of missing data, prioritises the most informative samples for annotation, and detects distribution shifts in out-of-distribution data to guide fine-tuning. By leveraging uncertainty, this thesis makes models more data-efficient, adaptive, and generalisable across diverse healthcare use cases.

The first contribution focuses on incorporating uncertainty-aware sequence-to-sequence predictions on sparse time series. These are often sparsely-sampled due to missed recordings, device malfunctions, or diverse sensing conditions. By enhancing uncertainty estimation in evidential deep learning and introducing metrics for assessing uncertainty estimations on sequence-to-sequence predictions, the proposed methodology provides a

reliable way to design sequence-to-sequence prediction models for real-world use cases.

Further, this thesis proceeds with a contribution addressing the bottleneck of unlabelled time series by automating machine learning workflows in a way that integrates human expertise and automated model refinement. Recognising the costly nature of labelling biosignals and the limited technical expertise of medical experts in model optimisation, the proposed approach facilitates the selection of highly-informative samples for annotation, dynamically tunes models during training, and maximises the use of unlabelled data. This method continuously refines the model with expanding data and human input, outperforms baselines and state-of-the-art methods, and reduces reliance on manual annotation and parameter tuning, thereby maximising the information gained through each annotation step.

Finally, this thesis also focuses on enhancing model adaptability under distribution shifts in healthcare time series, where datasets collected in one context often differ from those encountered in deployment. By leveraging uncertainty to recognise, quantify, and adapt to diverse and overlapping shifts, the proposed method improves the fine-tuning of pre-trained models under out-of-distribution conditions. This allows models to better handle multiple types and levels of distribution shifts, ensuring more reliable and effective adaptation, while avoiding overfitting and inefficient use of computational resources.

Overall, by addressing the limitations imposed by sparse data, unlabelled data, and distribution shifts, this thesis contributes to the design of more generalisable, data-efficient, and adaptive algorithms for healthcare time series analysis.

Acknowledgments

Reaching the stage of writing a PhD thesis is a unique opportunity, and one for which I am deeply grateful and will forever remember. However, this is something that I could not have achieved alone. It is thanks to the support of exceptional mentors, inspiring colleagues, close friends, and my ever-supportive family that I have made it this far. I will forever be thankful to each of them for their unwavering care and unconditional help.

First, I owe my deepest thanks to my PhD supervisor, Prof. Cecilia Mascolo, whose guidance has been invaluable throughout my studies. From the very beginning, her insightful feedback and steadfast support have helped shape both my research and personal development. The generosity of her time and expertise, combined with her constant encouragement, has made this journey not only intellectually rewarding but also profoundly meaningful. I feel incredibly fortunate to have had such a welcoming and encouraging supervisor, whose openness and dedication have made this path inspiring and fulfilling.

I would also like to extend my sincere thanks to the Mobile Systems group, including both current members and those who were part of the group when I joined. I felt welcomed from the very beginning, and our discussions helped me orient myself in the life and expectations of a researcher. I am deeply grateful to all my mentors and collaborators for their guidance. Ting, Lorena, George, and Yvonne, you have made my research endeavours infinitely easier through your support and continuous help. I have learned so much from you, and I couldn't have asked for better mentors. Similarly, I would like to thank the students I have supervised. Andres and Jenna, you taught me so much through our collaboration, and I'm very thankful for the opportunity to work together.

My heartfelt thanks also go to my labmates. Jake, Jing, Evelyn, Kayla, Adam, Yang, Zoey, Mathias, Tong, Ian, Young, Sudo, Erika, Hong, Abhirup, Qiang, and Gaia, not only did you help me explore new research directions, but you also became friends for life. Our days in the lab, the mystery around Secret Santa at our Xmas dinners, and our many other shared moments made my time in Cambridge truly unforgettable. I feel incredibly

lucky to have shared this journey with such a supportive and inspiring group of people.

Equally, my thanks go to my colleagues from the Sensors CDT cohort, who have not only been a source of academic inspiration but also wonderful companions throughout the years. Whether at group seminars, team-building sessions, or the memorable summer barbecue, these experiences have shaped my time in Cambridge and left me with fond memories. I am also deeply thankful to my dear friends from college. I will always treasure the memories of our time together, from those unforgettable punting trips along the River Cam to the dinners in the college dining hall that brightened even our busiest days. You have all been a constant source of motivation for me. I am also deeply thankful to my lifelong friends from UCL. I'll always cherish the moments we shared, from laughing our way through study sessions in the library, to celebrating the end of exam seasons. I'm certain that these experiences will stay with me long after my university years.

Of course, I also owe a special thanks to my examiners, Professors Katayoun Farrahi and Dong Ma, for their time, care, and thoughtful engagement with my thesis. I am especially grateful for their insightful suggestions and valuable feedback, which have meaningfully strengthened both the clarity and the quality of this work and have inspired me to pursue the highest standards in my future research.

A special shout-out also goes to all the academics who have taught me throughout the years, from my undergraduate studies all the way through to my doctoral path. I am certain that every piece of knowledge gained in my educational journey has positively impacted this thesis. In particular, I would like to acknowledge my MPhil supervisor, Prof. Neil Lawrence, who introduced me to the fascinating world of AI and Cambridge's academic culture, as well as Andrei, whose guidance and support during my MPhil days were invaluable. Their mentorship laid a crucial foundation for my studies, and I am deeply thankful for their help.

My heartfelt thanks also go to the administrative staff from the Department of Computer Science & Technology and the Sensors CDT, whose support enabled me to pursue my research, present my work at international conferences, and engage with the academic community. I am similarly grateful to Arm and the EPSRC for funding my PhD studies and making this work possible. Their combined contributions, whether academic, logistical, or financial, have made this journey far smoother than I could ever have imagined.

Last but not least, I owe my sincere gratitude to my parents and my entire family, who have always been there for me, prioritising my education and supporting me every step of the way. They have been my strongest supporters, and I am incredibly lucky to have them by my side throughout this long journey.

Thank you all for your encouragement, inspiration, and belief in me. This thesis would not have been possible without you, and I carry a deep sense of gratitude for each of you.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

Today, more than ever, Machine Learning (ML) models for time series analysis are increasingly popular in areas like healthcare, where they support personalised monitoring, diagnostics, and performance tracking [1]. The increasing availability of data from wearable sensors, mobile devices, and continuous monitoring technologies has opened unprecedented opportunities for real-world applications [2]. Specifically, sensor-based datasets have the potential to provide a more complete and continuous view of an individual's health and fitness, including data collected both in-the-wild and in hospital visits.

ML models have been widely used for things like wellbeing assessment and heart rate monitoring, but data from body-worn sensors have recently made them even more powerful through electrocardiograms (ECGs) for heartbeat categorisation and arrhythmia detection [3], through electroencephalograms (EEGs) for epileptic seizure recognition [4], through inertial measurement unit sensors (IMUs) for human activity recognition [5], and through continuous glucose monitoring (CGM) devices that help those suffering from diabetic disorders to detect and treat things like hypoglycaemia [6]. Such datasets comprise measurements and recordings of various functions and processes of the body, including vital signs, physical activities, and physiological responses. Simultaneously, ML algorithms have made remarkable progress in extracting meaningful insights and learning useful representations from such data. The convergence of rapidly growing medical data availability and increasingly powerful ML techniques has enabled a new generation of sensing-based applications that were previously unattainable. Together, these developments offer domain experts like healthcare professionals valuable insights to support more informed and effective decision-making. In practice, this means that models can assist in tasks such as early detection of cardiac abnormalities from ECG streams or predicting glucose fluctuations to support timely insulin administration, positively transforming patient health.

Yet, despite the potential of this area, the reality of using machine learning solutions

for healthcare time series remains highly challenging. Physiological datasets collected outside controlled environments are frequently sparse (i.e., contain missing observations), as sensors may occasionally fail and environmental factors may interrupt their operations [7]. For instance, a wearable glucose monitor may temporarily lose signal due to poor skin contact or motion artefacts during daily activities, resulting in gaps in the recorded data. Moreover, many of the available datasets are unlabelled or only partially labelled, as annotation requires costly expert supervision [8]. In real-world clinical settings, this corresponds to situations where only a small fraction of recordings are reviewed and annotated by specialists, leaving large volumes of potentially useful data unexploited. Further, models trained on one dataset often fail to generalise and struggle under out-of-distribution (OOD) conditions [9], which are caused by variations in populations, sensing devices, or collection protocols, thus limiting their reliability in deployment settings. For example, a model trained on data from hospital-grade ECG devices may perform poorly when applied to signals collected from consumer-grade wearables, or when deployed across different patient demographics.

These factors act as a bottleneck to the use of real-world time series for training deep learning models and are particularly critical in healthcare. An unreliable model may not only yield poor predictive performance but also lead to overconfident metrics. In safety-critical scenarios such as detecting arrhythmias or predicting hypoglycaemic events, such overconfidence can translate into missed alerts or inappropriate interventions, posing risks for patients. In this context, quantifying uncertainty plays a central role as a mechanism to enhance the ML training process itself [10]. Specifically, it can help address missing data points in sparse time series by providing information about the significance of their missingness, guide data annotation by prioritising the most informative samples for labelling, and signal when a distribution shift has occurred in out-of-distribution data to perform more carefully-guided fine-tuning. For instance, uncertainty estimates can highlight segments of a glucose time series where predictions are unreliable due to missing readings, prompting either cautious clinical interpretation or targeted monitoring. By leveraging uncertainty in these ways, models can become more data-efficient, adaptive, and robust, ultimately improving their ability to generalise across diverse use cases [11].

This thesis is motivated by these challenges and aims to develop machine learning methods that move beyond the narrow settings of conventional models. Specifically, it focuses on:

- Incorporating uncertainty-aware sequence-to-sequence prediction on sparse (i.e., with missing observations) healthcare time series data.
- Automating the ML pipeline, focusing on efficiently labelling highly-informative data while leveraging the inherent structure of unlabelled data, to reduce reliance on costly manual annotation.

- Enhancing model adaptability under distribution shifts using uncertainty information, to ensure reliable and effective fine-tuning of pre-trained models on OOD datasets.

To sum up, this thesis contributes to the design of novel and more generalisable algorithms for healthcare time series analysis through uncertainty-aware learning, addressing the limitations imposed by sparse and unlabelled data while also improving robustness to distribution shifts. It thus helps towards more reliable, generalisable, and effective machine learning models for real-world healthcare use cases.

1.2 Challenges and Research Questions

Healthcare time series often present challenges such as missing observations, missing labels, and distribution shifts between training and deployment environments. These limit the effectiveness of standard ML models, which typically assume regularly-spaced, fully-labelled data and stable data distributions. To help ML models in realising their potential, this thesis focuses on three key challenges:

Challenge 1: Modelling Sequence-to-Sequence Solutions for Sparse Time Series

Healthcare time series are rarely complete or evenly spaced. Specifically, it is often the case that gaps in the data arise due to missed sensor recordings, device malfunctions, or diverse user measurements with various sensors [7]. In practice, this can correspond to irregular glucose measurements when a CGM device or other physiological monitor temporarily loses contact with the patient. Most machine learning models assume full data availability, which leads to performance degradation and unreliable predictions when applied to sparse streams. Uncertainty estimation can help models to perform better in the presence of sparsity, but while it has been explored in single-point prediction tasks, its integration into sequence-to-sequence models for sparse time series remains underdeveloped [11]. Sequence-to-sequence predictions involve mapping an input sequence to an output sequence, which is important because many real-world problems depend on capturing relationships across entire sequences rather than isolated points. For example, predicting the future trajectory of a patient's glucose levels requires understanding temporal patterns over extended periods rather than single observations. This gap is particularly important in healthcare time series, where the information at missing time steps can prove crucial for sequence-to-sequence predictions [7].

Challenge 2: Automating ML Workflows for Unlabelled Time Series

Despite the ever-increasing availability of sensor-generated datasets, a major bottleneck to the further expansion of healthcare ML is the presence of unlabelled samples in them [2]. Unlike other domains where annotation can be more straightforward, labelling biosignals requires medical expertise, which is both costly and time-consuming. For instance, annotating ECG recordings for arrhythmia detection often requires cardiologists to manually review long signal traces, making large-scale labelling impractical. Current active learning strategies partially alleviate this but they treat annotation, hyperparameter tuning, and model training as decoupled processes [8, 12]. This results in inefficient use of data and annotation budgets, as the models are rarely optimised for the specific dataset, and they do not dynamically adapt to the newly-gained knowledge about the labels of its various samples at annotation time. Moreover, semi-supervised learning makes the most of unlabelled samples with limited human input [13, 14], but this often comes at the expense of reduced accuracy as the knowledge that can be gained by medical professionals is not effectively passed to the model. Therefore, automated and adaptive workflows that integrate human expertise, unlabelled data, and model optimisation seamlessly in healthcare time series are still lacking.

Challenge 3: Ensuring Robustness under Distribution Shifts in Healthcare Time Series

Data samples collected in one context often differ from those encountered in deployment, and variations in hospitals, sensors, and patient populations, can cause distribution shifts that lead models trained on data with even slightly different characteristics to underperform in practice. In real-world deployment, this may arise when a model trained on data from one hospital is applied to another with different equipment or patient demographics, or when transitioning from controlled clinical environments to at-home monitoring scenarios. This is important as some time series datasets, such as those from ECGs, pose even more significant complexities in labelling, leading to the use of solutions that rely on human-in-the-loop annotation less suitable for them [15, 16]. The emergence of pretraining methods helps alleviate this, but their performance often deteriorates and they tend to overfit and not generalise due to out-of-distribution (OOD) effects [9]. This is especially the case in more complex scenarios involving multiple types, levels, and combinations of shifts. Existing one-size-fits-all methods tend to ignore the granularity or severity of the shifts in ECG data [17, 18, 19]. For example, adapting to a large dataset differs from adapting to a small one with a new task, yet both are often treated similarly, leading to suboptimal model performance and inefficient use of computational resources due to mismatched adaptation strategies [20, 9]. Therefore, it is important to quantify distribution shift severity, and use this information alongside further settings to improve

the fine-tuning of pre-trained models as much as possible.

1.3 Research Questions

Motivated by the challenges discussed above, this thesis aims to design novel algorithms that focus on the development of more generalisable models for real-world healthcare time series applications. This is pursued by effectively incorporating sparse data, better utilising unlabelled or limited labelled datasets, and improving model robustness to distribution shifts. In particular, these efforts are directed toward answering the following research questions:

- **Research Question 1:** How can we design sequence-to-sequence models that effectively capture temporal dependencies in sparse healthcare time series while providing reliable and informative uncertainty estimates?
- **Research Question 2:** How can we develop automated learning frameworks that efficiently exploit unlabelled and partially labelled healthcare time series, while minimising human annotation effort and reducing dependence on expert supervision?
- **Research Question 3:** How can we design models that generalise across heterogeneous time series distributions and remain effective under complex and overlapping distribution shifts in healthcare deployment settings?

1.4 Contributions and Chapter Outline

This thesis will start with an overview of the background and existing works in sparse, unlabelled, and out-of-distribution biosignals in Chapter 2, before presenting the three main contributions that address the research questions posed in the previous section. These contributions are summarised below:

1.4.1 Contribution 1: Sparse Sequence-to-Sequence Uncertainty Estimation in Evidential Deep Learning

Machine Learning models typically assume that time series are regularly spaced, however this is often unrealistic in healthcare, where missing data recordings are common. In this context, uncertainty estimates play a pivotal role, as they can enable confident and non-confident predictions to be distinguished. This is the highlight of Chapter 3, where we propose SQUIREDL: a novel uncertainty-aware sequence-to-sequence prediction method for sparse healthcare time series. Specifically, we enhance the state-of-the-art evidential regression framework, widely used for uncertainty estimation, to handle missing data.

Following data imputation with an Akima spline-based method, we modify the loss function of evidential regression by assigning different weights to imputed and observed data points, to offer more reliable uncertainty estimates. Additionally, we examine a variety of metrics for assessing the success of uncertainty estimations on sequence-to-sequence predictions, providing a reliable way to evaluate the models in a medical setting. Our proposal is demonstrated in two clinical applications. In continuous glucose monitoring, we use sequence-to-sequence prediction to obtain the hypoglycaemia risk from glucose sensor readings. Our approach captures the ground truth risk values 30% more accurately, bringing consistent improvements in both uncertainty-aware and accuracy-based metrics. Similarly, in COVID-19 hospital admissions data, we achieve a 22% improvement in the accuracy of uncertainty-aware predictions, enabling better resource planning.

1.4.2 Contribution 2: Streamlined Adaptive Learning for Unlabelled Sensors Time Series

Further to the sparsity aspects examined in Contribution 1 above, sensor-generated time series also present challenges in labelling due to their sequential nature, which requires consideration of context and temporal dependencies. Recognising the costly nature of data labelling and that domain experts may have limited technical expertise in model optimisation, in Chapter 4 we introduce SALTS: an approach to automate machine learning model training for medical time series, enhancing analysis efficiency. This first operates at the data input level via adaptive data acquisition, facilitating the selection of highly-informative samples for labelling. Further, it works at the model level, through dynamic model refinement to optimise the model on-the-fly by progressively exploring the possible hyperparameter options and choosing the best combination at each acquisition step, and through an automatic learning phase to maximise the usage of any unlabelled samples. This results in a robust learning strategy that continuously refines the model with expanding data and human expertise. Demonstrated on EEG, ECG, and IMU health signal classification, our method outperforms baselines and the current state-of-the-art, while reducing reliance on human input for model tuning. SALTS enhances the applicability of machine learning to healthcare time series, maximising the information gained through each human annotation step in an automated way.

1.4.3 Contribution 3: Uncertainty-Aware Fine-Tuning for Out-of-Distribution ECG Time Series Models

Data samples collected in one context often differ from those encountered in deployment, leading to distribution shifts that lead models trained on datasets with even slightly different characteristics to underperform. Recognising that existing solutions are often not

effective in recognising, quantifying, and adapting to diverse and overlapping shifts in electrocardiogram (ECG) time series, in Chapter 5 we propose ADAPTOOD: a framework that leverages data uncertainty to quantify distribution shift severity, and uses this information to improve the fine-tuning process. To quantify the severity we leverage the differences between the pre-training and OOD data distributions. This form of uncertainty naturally correlates with the fine-grained OOD severity, making it a practical guidance signal for adaptation. Building on this, we use a combination of low-rank model updates and adaptive hyperparameter optimisation to further support the adaption mechanism. ADAPTOOD effectively handles varying levels of distribution shifts, enabling robust, efficient, and accurate model updates across diverse OOD scenarios. We evaluate our method across real-world ECG OOD use cases and demonstrate that it delivers up to 7% higher accuracy and 12.9% higher precision compared to existing methods.

1.5 Research Outputs

The research described in this dissertation has led to several works that have been published, presented, or submitted to various conferences, journals, and symposiums. These are highlighted below.

Works Related to this Thesis

This list highlights my first-author papers, whose contributions are directly relevant to this thesis.

[21] "SQUIREDL: Sparse Sequence-to-Sequence Uncertainty Estimation in Evidential Deep Learning"
S. Vavaroutas, T. Dang, E. Rocheteau, and C. Mascolo. *ACM Transactions on Computing for Healthcare*, vol. 6, no. 3, 2025.

[22] "SALTS: Streamlined Adaptive Learning for Sensors Time Series"
S. Vavaroutas, G. Rizos, and C. Mascolo. In *Proceedings of the 47th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2025.

[23] "ADAPTOOD: Uncertainty-Aware Fine-Tuning for Out-of-Distribution ECG Time Series Models"
S. Vavaroutas, Y. Wu, A. Etemad, and C. Mascolo.

Other Works

Additionally, I contributed to other works in the area of machine learning for biosignals, which influenced the ideas and development of this thesis and are listed below.

[24] "Uncertainty Estimation with Data Augmentation for Active Learning Tasks on Health Data"
S. Vavaroutas, L. Qendro, and C. Mascolo. In *Proceedings of the 45th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2023.

[25] "Uncertainty Estimation for Sequence-to-Sequence Regression on Sparse Time Series"
S. Vavaroutas, T. Dang, E. Rocheteau, and C. Mascolo. *Presented at the 5th UK Mobile, Wearable and Ubiquitous Systems Research Symposium (MobiUK)*, 2023.

[26] "Benchmarking Foundation Models on Out-of-Distribution Wearable Biosignals"
A. Alvarez Olmo, S. Vavaroutas, Y. Wu, and C. Mascolo. *Presented at the 7th UK Mobile, Wearable and Ubiquitous Systems Research Symposium (MobiUK)*, 2025.

Chapter 2

Related Work

In the previous chapter, we examined the potential and challenges of developing and deploying ML models for the analysis of healthcare time series, as well as the crucial role of uncertainty-aware learning in such scenarios. In this chapter, we provide a background on the use of biosignals for healthcare tasks (Section 2.1), and cover the machine learning theory used as a foundation for this work (Section 2.2). We then proceed with reviewing the relevant literature on sparse time series analysis (Section 2.4), on automated solutions for unlabelled time series model development (Section 2.5), and on the importance and solutions around handling distribution shifts in real-world settings (Section 2.6).

2.1 Role and Significance of Biosignals

Biosignals are quantitative measures of physiological processes such as heart activity through electrocardiograms (ECG), brainwaves through electroencephalograms (EEGs), human activity through inertial measurement unit sensors (IMUs), and glucose levels through continuous glucose monitoring (CGM) sensors. These have long played a foundational role in clinical diagnosis and monitoring [1, 2], enabling robust and effective healthcare support. For example, ECG signals are routinely used in hospitals to detect cardiac arrhythmias, while CGM sensors enable individuals with diabetes to continuously track glucose levels and make timely decisions about insulin intake in daily life.

Traditionally, acquiring such data required the use of specialised equipment in clinical settings, limiting their applicability to more frequent and continuous snapshots of a patient's condition [27]. Medical staff interpreted these measurements based on established patterns, often relying heavily on manual analysis and expert judgment. While effective in controlled settings, this approach posed challenges in terms of scalability, continuous monitoring, and wide reach of the relevant technologies in real-world deployments. In practice, this meant that conditions with intermittent symptoms could go undetected between clinical visits, and long-term trends in a patient's health were difficult to capture

reliably.

While the history of the ECG has been complicated and unorganised, it is noteworthy that more than 100 years have passed since the first ECG of a human being was recorded [28]. Since then, the clinical use of ECG has been well studied, a wealth of information about ECG patterns and arrhythmias has been widely available for many years [28], and ECGs have come to work by measuring the electrical activity of the heart through electrodes placed on the skin, which detect the flow of ions that generate the heart's electrical impulses. These measurements form signals where subtle variations in waveform shape or timing can indicate clinically significant conditions, such as atrial fibrillation or myocardial infarction. Similarly, there have been several decades of modification and advancement around the use of the EEG, which relies on electrodes and coin-sized metal disks connected to thin wires to record electrical impulses sent out through brain cells [29]. The insights and discoveries found using an EEG are significant, enabling medical professionals to identify obscure frequencies or patterns in the brain's activity [29]. For instance, abnormal rhythmic patterns in EEG recordings are commonly used to detect and monitor epileptic seizures in both clinical and ambulatory settings.

Although the evolution of wearable IMUs has been more recent, they have also switched from their use in supervised laboratory and clinical settings to their application for long-term monitoring of human movement in unsupervised real-world settings [30, 31]. Mobile and wearable sensors offer a versatile and accessible platform for IMU-based health applications [32]. In real-world scenarios, these are embedded in smartphones or fitness trackers to capture daily activities such as walking, sitting, or exercising, supporting applications ranging from activity monitoring to fall detection in elderly populations. With regards to CGM sensors, the release of commercial and portable blood glucose meters in the late 20th century transformed glucose monitoring [33]. These sensors transformed diabetes management by providing real-time glucose readings using electrodes paired with redox enzymes in glucose oxidase, which generate an electric current proportional to glucose concentration that is then converted into a readable glucose measurement [33]. This enables continuous tracking of glucose fluctuations throughout the day, allowing users to observe how factors such as meals, physical activity, or stress impact their glucose levels.

Despite the significant development of such technologies, it is the emergence of pervasive platforms [34], wearable sensors [35], and mobile technologies [36] that transformed the landscape by enabling healthcare data to be collected outside of clinical settings [37]. Devices such as smartwatches, fitness trackers, and other wearable monitors [38] now routinely collect continuous time series streams of physiological data, providing a more dynamic and contextually rich picture of an individual's health. For example, they may continuously monitor heart rate variability during sleep and daily activities, offering insights into stress levels and cardiovascular health over extended periods. However, this influx of data has also posed new challenges: how to manage missing data points and

variability, how to annotate the samples to extract meaningful insights, and how to make accurate predictions under varying contexts.

To this purpose, machine learning solutions came to revolutionise biosignal analysis. By learning patterns directly found in the data, machine learning algorithms help to identify complex relationships between physiological signals and health outcomes that would be difficult, costly, and time-consuming to model manually [27]. For instance, ML models can automatically learn to distinguish between normal and abnormal heart rhythms from raw ECG signals, or detect early signs of deterioration in patient health from other sensor streams. Early ML solutions proved effective for limited tasks from carefully-processed datasets, but the rise of deep learning has brought a huge leap in capabilities, enabling feature extraction from raw biosignals [37]. However, these models now need to be improved in terms of their adaptability: they need to better handle sparse data, learn from limited labelled samples, and generalise across populations and acquisition sensors/devices. These factors should be taken into account to help AI-based health diagnoses become feasible in everyday settings, facilitating more personalised healthcare solutions.

2.2 Neural Networks and Their Use in Time Series

Despite current challenges regarding data sparsity, labelling, and applicability to other tasks, the fusion of biosignal monitoring with machine learning marks a significant shift from reactive care to continuous data-driven healthcare intelligence. Neural networks have played a paramount role in this.

Neural networks are computational models designed to recognise patterns and make predictions based on the data fed to them, and they are inspired by the biological structure of the brain [39]. Typically, they consist of layers of nodes called neurons, with each of those performing a mathematical operation and passing its output to the next layer. The most basic form of a neural network is a single-layer network called the perceptron, while more advanced models, such as deep neural networks, contain multiple layers that allow them to learn hierarchical representations of the data [39]. Their main advantage is that they can model more complex and non-linear relationships between input and output data, making them widely applicable across domains ranging from image recognition to time series classification tasks [40].

The learning process of neural network-based models is typically governed by the backpropagation algorithm, which allows them to iteratively adjust their internal weights by minimising a loss function so that they learn to approximate the mapping between the input and output data [41]. While traditional fully connected architectures like multilayer perceptrons (MLPs) are capable of capturing intricate relationships in the data [41], they are not always the most efficient for tasks involving sequential or temporal data like time

series that are the focus of this thesis, where relationships between observations evolve over time.

For time series data, Convolutional Neural Networks (CNNs) have proven to be particularly effective, as they have been widely applied to one-dimensional temporal data and they have been shown to excel at recognising local and shift-invariant patterns along the time axis [40]. This is particularly important in time series analysis, as meaningful patterns such as peaks, oscillations, or anomalies often repeat at different points in time but share similar characteristics. CNNs operate by applying small filters, which are called kernels, that slide over the input data, learning local dependencies temporal features [39]. Unlike Recurrent Neural Networks (RNNs), which process inputs sequentially and rely on maintaining internal states across time steps [42], CNNs perform operations in parallel, which makes them faster to train and more scalable when working with large datasets.

Additionally, in many real-world biosignals such as ECG or EEG recordings, important patterns can appear at different points in time but still represent the same critical underlying medical event. CNNs can detect these temporal structures regardless of their exact position in the sequence, making them robust to variations in timing [39]. For instance, they can effectively identify heart abnormalities in ECG signals or event-related potentials in EEG signals, even if these occur at different times, while they can also learn hierarchical representations through multiple layers of convolution and pooling. Their early layers tend to capture simple features such as short-term fluctuations or periodic components, while deeper layers extract more abstract patterns representing long-term dependencies or more complex temporal trends [43]. Further, pooling layers are used to reduce the dimensionality of the feature maps, improving efficiency and helping the model to better generalise to unseen data. This layered structure enables CNNs to effectively capture both local and global temporal dynamics [43].

Last but not least, CNNs can often learn directly from raw time series biosignals, without the need for extensive feature engineering [37]. This is particularly valuable in healthcare and wearable sensing applications, where features might not generalise well across different populations, sensor types, or recording conditions. By learning relevant features from the raw input signals, CNNs can adapt to diverse data sources and tasks with less pre-processing than other types of networks. As such, CNNs are one of the most powerful and efficient architectures for time series analysis, striking an effective balance between performance, efficiency, and representational power, and they are thus used extensively in this thesis.

2.3 Uncertainty in Time Series

Uncertainty quantification has become imperative in domains such as healthcare, where predictions need to not only be accurate but also reliable. Traditional models often focus

on predictive accuracy, but this is often insufficient in real-world applications where understanding the confidence of predictions is equally important. A wide range of approaches have been proposed to quantify this confidence through uncertainty in deep learning models for time series settings, each with distinct assumptions, computational trade-offs, and applicability across different settings.

2.3.1 Uncertainty as a Measure of Time Series Model Confidence

In terms of techniques for uncertainty estimation, Monte Carlo (MC) dropout [44] is one of the most widely used approaches. It approximates Bayesian inference by performing multiple stochastic forward passes through a neural network with dropout enabled at test-time, generating a distribution over predictions instead of absolute values. It has been successfully applied to time series forecasting and uncertainty quantification [45], and it allows models to capture epistemic uncertainty without explicitly modelling posterior distributions over weights. However, it requires a large number of samples to obtain stable uncertainty estimates, which significantly increases computational cost, relying on extensive sampling during inference. Additionally, the quality of its uncertainty estimates depends on the choice of the dropout rate and may degrade in complex temporal settings, despite recent extensions to uncertainty-aware interpretability [46].

Bayesian techniques have also been used for uncertainty estimation [47]. They aim to model uncertainty by placing probability distributions over network weights and performing posterior inference given the data [48]. This can capture both epistemic and aleatoric uncertainty and has been explored in benchmarking toolboxes for industrial time series forecasting [49]. However, exact inference is intractable for deep architectures, and approximate methods such as variational inference are often difficult to scale and require careful tuning of priors and approximations, limiting their applicability.

Further, deep ensembles are another way of capturing time series uncertainty [50] and they rely on training multiple independently initialised models. They estimate uncertainty by aggregating predictions from diverse models and have been further enhanced through optimisation strategies such as Bayesian-optimised ensembles [51]. They have been shown to provide strong empirical performance and robust uncertainty estimates, but the need to train and store multiple models leads to increased computational requirements, and they may still struggle to capture uncertainty when model diversity is insufficient or not optimised.

2.3.2 Conformal Predictions

More recently, conformal predictions have emerged as another framework for uncertainty quantification, providing distribution-free prediction intervals for ML models [52]. Con-

formal predictions operate at a user-specified significance level, and, unlike Bayesian or sampling-based approaches, they operate as model-agnostic wrappers [53]. Specifically, they leverage nonconformity scores computed on a previously available calibration dataset to construct prediction regions with guaranteed coverage, assuming exchangeability of the data. An exchangeable data sequence can be arbitrarily long, and the term exchangeable means that its joint probability distribution should not change when the positions of data points in the sequence are altered. This allows conformal predictions to provide statistically valid uncertainty estimates without modifications to the predictive model or strong assumptions about the underlying data distributions [53].

Of note, conformal predictions are able to provide uncertainty estimates for individual predictions. This is imperative in healthcare applications, where decisions often have to be made per patient rather than for an entire population, and where global metrics such as accuracy fail to capture prediction reliability [54]. Recent work has demonstrated that conformal predictions can enhance decision support systems by identifying unreliable predictions [55], while their use has also been explored in dynamic biological systems as an alternative to traditional Bayesian approaches, offering non-asymptotic guarantees without relying on parametric assumptions [56].

Despite the above, conformal predictions can still face challenges when applied to real-world time series. Their validity relies on the assumption of exchangeability, as discussed above, which is often violated in sequences with temporal dependencies. Recent work has proposed extensions like adaptive conformal inference to address such temporal dependencies [57]. However, these approaches typically introduce additional complexity or trade-offs between efficiency and validity, while conformal predictions primarily produce intervals that do not capture the decomposition of uncertainty into aleatoric and epistemic components. Aleatoric uncertainty stems from inherent randomness in the data and cannot be reduced with more data, while epistemic uncertainty is caused due to lack of knowledge and can be reduced by feeding more training data to the model. Knowing the uncertainty type is imperative in mitigating it, so the fact that conformal predictions do not help to identify the source of the uncertainty may limit their applicability.

2.3.3 Evidential Deep Learning for Uncertainty Estimation

In contrast to sampling-based approaches, methods based on deep evidential regression offer a compelling alternative. To this purpose, evidential deep learning incorporates an additional layer to estimate both a continuous target and the associated evidence directly within a single model [58]. This direct estimation of evidence allows it to differentiate between aleatoric and epistemic uncertainty more precisely, leading to better-calibrated and more informative uncertainty measures. This can be used to regularise the model when the predicted evidence is not aligned with the correct output. Thus, deep evi-

dential regression achieves uncertainty estimation through a deterministic network [59], significantly reducing computational complexity.

Deep evidential regression [58], an extension of evidential deep learning [59] for regression tasks, stands out among alternatives as it formulates learning as an evidence-acquisition process and imposes fewer assumptions, offering improved calibration. In particular, it models regression targets using a normal inverse-gamma distribution, allowing the network to simultaneously predict both the mean and variance of the target as well as the uncertainty in those estimates [58]. By penalising inconsistent or unjustified evidence during training, the model is encouraged to provide calibrated uncertainty estimates that reflect its confidence in predictions. This approach captures both aleatoric and epistemic uncertainty without relying on sampling or ensembles, making it efficient and scalable. Although evidential deep learning and its extensions have shown superior performance in regression tasks [58] and in tasks with imbalanced data [60], their capacity in real-world sequence-to-sequence predictions remains an underexplored area.

2.3.4 Uncertainty under Sparse Data

While uncertainty estimation methods are often studied independently, their behaviour is closely tied to characteristics of time series data, such as missingness. Yet, existing uncertainty-aware approaches rarely account for the fact that uncertainty estimations are directly affected by the extent of data missingness. This highlights the need for solutions that explicitly distinguish between observed and imputed values during training, while preserving temporal dependencies in sequence-to-sequence settings.

The methodology of Neural Ordinary Differential Equations (NODEs) [61] can model continuous-time dynamics by parametrising the derivative of the hidden state through a neural network in cases of sparsity. They achieve this by modelling the continuous-time evolution of a latent state, only evaluating the loss function at the actual observation times and simply integrating the learned dynamics forward without defining values at missing timestamps [62, 63]. Yet, the selection of solvers and initial conditions, and the varying ODE structures for different types of time series remain significant challenges. It has also been noted that achieving real-time performance utilising a Long Short-Term Memory (LSTM) network [64] to capture long-term dependencies with NODEs is more challenging for time series. On the contrary, integrating neural architectures with evidential regression approaches enhances modelling flexibility with minimal computational complexity, while also enabling the separation of epistemic and aleatoric uncertainties.

Other streams of work have developed different methods for uncertainty estimation in medical time series [47, 65], but these still neglect the temporal dynamics of the series or the uncertainty in sequence-to-sequence tasks. Such solutions may also fail when used for sparse time series, due to the assumption of evenly-spaced sequences that misguides the

model in learning the missingness prevalent in the data. Moreover, recent survey papers have shown that uncertainty estimation can help in informing ML model development and deployment in healthcare [11], showcasing their importance in supporting clinical decision-making.

2.3.5 Predictive Uncertainty for Targeted Data Acquisition

Beyond its role in modelling confidence, predictive uncertainty has also been explored as a proxy for selecting informative unlabelled instances. This is particularly helpful for classification tasks, where uncertainty sampling [66] is often used to select samples for which the model exhibits low confidence, typically quantified through posterior class probabilities. In its simplest form, this can be expressed as $U(x) = 1 - P(\hat{x} \mid x)$, and this strategy has been widely adopted due to its simplicity and effectiveness, particularly in scenarios where labelling costs are high [8]. By focusing on acquiring data labels for samples near the current decision boundary, uncertainty-based methods aim to refine the model where it is most likely to make errors, thus accelerating convergence [67].

Beyond classical settings, uncertainty estimates have also been leveraged to guide the use of unlabelled data. In particular, pseudo-labelling techniques rely on model confidence to determine which predictions are sufficiently reliable to be treated as ground truth [12]. In this case, uncertainty acts as a filtering mechanism: low-uncertainty samples are incorporated into the training set, while highly-ambiguous samples are deferred to avoid propagating noise. These ideas further highlight the broader role of uncertainty beyond prediction, and the connection between uncertainty and data selection is explored in more detail in Sections 2.5.2 and 2.5.3.

2.3.6 Uncertainty for Distribution Shift Identification

Uncertainty has also become central to detecting distribution shifts. In these cases, uncertainty can act as a proxy for the divergence between a source distribution D_s and a target distribution D_t, where increased uncertainty reflects reduced overlap and thus a higher likelihood of distributional shift [68]. This has motivated a wide range of approaches that leverage uncertainty signals to identify when inputs deviate from the training distribution.

Both distance-based and density-based uncertainty measures provide a complementary perspective by directly quantifying discrepancies between distributions in feature space. These approaches are appealing as they can be computed prior to model adaptation, enabling proactive detection of potential distribution shifts. Recent work has also explored uncertainty estimates through improved representations and training objectives, for example by encouraging diversity in the learned weight space or incorporating entropy-aware formulations to better capture epistemic uncertainty [69]. Additionally, domain-specific studies highlight that the effectiveness of uncertainty-based shift detection is closely tied

to data structure and modality, with tailored approaches showing improved performance in high-dimensional settings [70, 11]. As uncertainty is key in handling distribution shifts, its use in model adaptation strategies is also discussed in Section 2.6.5.

2.4 Sparsity in Time Series

Sensor-based time series often include gaps caused by missed readings, device failures, or measurement errors from different sensors [7]. However, most ML models assume that datasets are sampled without missing observations, which can reduce performance and yield unreliable predictions when they are used on such sparse data. This issue is particularly pronounced in real-world domains such as healthcare and environmental monitoring, where data collection is often subject to hardware limitations and transmission errors [71].

Missing data not only reduces the effective size of training datasets but can also distort temporal dynamics, making it difficult for models to distinguish between true signal patterns and artefacts introduced by sparsity. As highlighted in recent surveys on climate time series [72], missingness is a pervasive challenge across domains, with statistical imputation methods such as interpolation, regression, and PCA being widely used, despite their limitations in capturing complex temporal dependencies when used in isolation.

2.4.1 Imputation and Sparsity-Aware Modelling

A common strategy for handling sparsity is to impute missing values prior to model training. Imputation-based approaches are popular due to their simplicity and efficiency, particularly in settings with limited computational resources. Among these, spline-based methods offer improved flexibility over linear interpolation by better adapting to intervals of data availability and missingness. Akima spline interpolation [73] constructs piecewise cubic polynomials that emphasise local variations, enabling more accurate reconstruction of abrupt changes in time series. This is especially beneficial in healthcare applications, where sudden physiological events manifest as sharp temporal fluctuations [74].

Beyond preprocessing, prior work has also explored incorporating sparsity directly into model design. For instance, recent work leverages a cross-period sparse forecasting mechanism that reduces model complexity while preserving long-term temporal structure [75], demonstrating that sparsity-aware representations can improve both efficiency and robustness. Similarly, sparsity-inducing methods such as the sparsity-ranked lasso [76] introduce structured regularisation to prioritise simpler temporal patterns, enabling effective modelling even in high-dimensional and partially observed settings. Beyond statistical approaches, deep learning models have also been extended to account for missingness patterns. Methods such as GRU-based architectures and sequence-to-sequence frameworks

have been adapted to handle sparse data in domain-specific applications like water quality prediction [77]. These models often incorporate attention mechanisms or feature interaction modules to mitigate the effects of sparsity, although they still typically rely on implicit handling of missing values, without modelling the reliability of imputed observations. However, existing approaches rarely consider how the imputation process itself affects downstream model confidence, leaving a gap in jointly modelling sparsity and uncertainty in sequence modelling settings.

2.4.2 Temporal Dependencies and Sequence-to-Sequence Modelling under Sparsity

Sequence-to-sequence models aim to produce an output sequence corresponding to a given input sequence [78]. While handling sparsity has been widely studied in single-point classification or regression tasks, its application in sequence-to-sequence settings remains underexplored.

Unlike traditional single-point classification or regression, the success of sequence-to-sequence models depends heavily on learning the temporal dependencies within data sequences. While well-established in domains such as language translation and voice-enabled commands [79], the consideration of temporal dependencies for sparse data remains largely unexplored, particularly in healthcare applications. The presence of gaps and inconsistencies in health data due to missed recordings or sensing errors [71] introduces additional complexity for producing reliable sequence-to-sequence predictions.

Towards addressing these challenges, recent advances in sequence modelling offer better ways to capture long-range dependencies in time series with missing data. For instance, diffusion-based approaches have been proposed for time series imputation, leveraging multi-scale architectures to model both local and global temporal patterns [80]. These methods address issues such as boundary inconsistencies between observed and imputed regions while improving the modelling of long-term dependencies. Similarly, hybrid architectures that combine transformer-based encoders with recurrent decoders have shown improved performance in multi-step forecasting tasks by integrating heterogeneous temporal features [81]. However, both types of approaches primarily focus on predictive accuracy for future forecasting and do not explicitly account for how missing or imputed values affect the reliability of sequence-to-sequence predictions.

Further, existing work for health time series mainly focuses on improving accuracy, overlooking the aspect uncertainty [82]. Despite some works analysing the uncertainty in irregular time series [83], these target single sequences of data, and have not examined sparsity in sequence-to-sequence modelling. Such time series pose a key challenge for uncertainty estimation. That is because their irregularities violate the modelling assumptions [71], making it difficult for traditional models, like those based on a recurrent

neural network [84], to grasp the temporal dynamics of the data. Additionally, despite current literature having shown a number of approaches to model sparse time series such as GRU-D [85] or an attention mechanism tailored to such time series [7], none of them is equipped with uncertainty estimation capabilities, nor do they provide a feasible way of integration with other approaches that are used for uncertainty estimation.

2.4.3 Summary

Overall, while prior research exists in irregular time series modelling, there remains a significant gap in integrating such approaches with uncertainty estimation for enhancing their robustness and effectiveness. Current studies rarely address uncertainty estimation within sequence-to-sequence frameworks for irregularly sampled time series, limiting their applicability to real-world sparse data. Therefore, it is imperative to address this by developing unified solutions that can simultaneously capture temporal dependencies and handle missing data through well-calibrated uncertainty estimates for sequence-to-sequence predictions.

2.5 Unlabelled Data in Time Series

Nowadays, the expanding availability of applications around deep learning in healthcare means that more and more datasets are needed for training the relevant models [86]. Despite the abundance of such data harvested from body-worn sensors, the key issue remaining is that the extraction of reliable labels for them is a challenging, manual, and costly operation. Additionally, the design and deployment of any ML model requires expert ML knowledge, as tasks like hyperparameter (HP) selection are non-trivial for someone with little ML background [87]. Unfortunately, domain experts like doctors, nurses and practitioners often lack both the time for labelling a sufficient number of data samples to feed in a model, and the machine learning expertise for designing the said model [88]. This limits the performance of deep learning models, which are inevitably more effective when they integrate the clinical expertise of medical professionals with the technical rigour of a well-designed system, rather than ignoring either [89].

To this purpose, human-in-the-loop strategies are essential for making the most of a human annotation budget. Prior work has explored these through Active Learning (AL) [90, 91], however such approaches often assume a fixed hyperparameter selection occurring before the data labelling, or a tuning step that remains static throughout the acquisition process: both result in suboptimal performance due to their inability to adapt hyperparameters during the acquisition. Additionally, an approach that aims to utilise information from unlabelled data, that of semi-supervised learning, often does not fully leverage domain expertise, overlooking the knowledge that medical experts can intro-

duce, [12] resulting sometimes in lower accuracy. Active semi-supervised learning [92] offers a step towards addressing this, though the need for manually tuning model hyperparameters at each data acquisition introduces complexity in addition to the laborious data labelling [2]. Automated methods for leveraging deep learning with minimal human intervention pose unique challenges, and this is exacerbated for healthcare time series due to their temporal nature.

2.5.1 Hyperparameter Optimisation

In model design, the careful and informed selection of hyperparameters is paramount, and can significantly help in improving performance. Hyperparameter optimisation determines the set of hyperparameters that yields an optimal model which minimises a predefined objective function on a given data set, such as validation loss, making this process critical for achieving the best results. [93]. Classical approaches often employ Bayesian optimisation, a widely used strategy that iteratively leverages information from prior evaluations to guide the exploration of the hyperparameter space [94]. Unlike less guided methods such as random or grid search, which may redundantly explore multiple local optima, Bayesian optimisation targets promising regions of the search space, yielding more consistent performance [94]. Recent surveys highlight that hyperparameter optimisation remains central to deep learning, affecting convergence speed, generalisation, and adaptation to certain tasks [95, 96].

In healthcare time series applications, automated hyperparameter tuning is even more critical, as domain experts may have limited machine learning expertise and the model design process adds further complexity. To this purpose, automated hyperparameter optimisation, a key aspect of AutoML [13], can systematically select optimal model configurations for a given dataset and task [97]. The use of AutoML has been explored for time series [98] and for healthcare use cases [99], yet the hyperparameter tuning of the active learning process remains underexplored, limiting the ability of models to adapt dynamically to newly labelled samples. Automating stages that otherwise require human intervention, such as hyperparameter tuning, is therefore essential for robust model development in domains with costly and complex annotation requirements.

2.5.2 Active Learning

Advances in active learning literature enable algorithms to achieve higher levels of accuracy given a constrained annotation budget, by iteratively selecting the data points that need to be labelled [8]. Prior work in biosignal classification [100], medical data analysis [101], and human activity recognition [91] showcases its efficacy in reducing annotation costs. Most approaches rely on uncertainty-based query strategies to identify informative samples, with demonstrated effectiveness in domains such as wearable sensing [100]. Sys-

tematic evaluations of active learning methods in healthcare [102] further confirm their practical benefits, although they are typically implemented within fixed model configurations during annotation, where the model is trained once and only incrementally updated with new labels without revisiting its hyperparameter space. This limits the ability of the model to adapt to newly-acquired data and evolving task complexity throughout the active learning process.

Prior literature looking at boosting the functionality of active learning systems with hyperparameter tuning capabilities is similarly limited. The most relevant study is Auto-DAL [103], which is targeted at a distributed computing setup, and involves partitioning data across machines, building independent models, and independently querying for labels based on partial views of the data, thus departing from the generality of the active learning task. It also uses a costly regularisation that is based on a radial basis function kernel, whose hyperparameters are iteratively tuned using optimisation methods [104, 105]. Instead, Bayesian optimisation approaches [94] can be more effective and efficient, while flexible hyperparameter optimisation methods [93] allow tuning of discrete design choices such as activation functions and optimisers. More importantly, many existing methods often treat hyperparameter optimisation as decoupled from the active learning loop, preventing adaptation to newly acquired labelled data.

On a related research contribution, collaborative filtering-based model selection methods [106] propose active strategies for time-constrained hyperparameter tuning, but assume fully labelled datasets and therefore do not operate within an active learning setting. Similarly, application-driven active learning frameworks in healthcare [107] and industrial time series [108] demonstrate strong performance gains, yet still rely on static or pre-configured model settings. The lack of mechanisms for smartly and continuously refining the model with expanding data and expertise as new labels are acquired, limits the effectiveness of solutions that rely solely on active learning for their operation, particularly in healthcare time series where annotation is costly and task complexity can vary significantly.

2.5.3 Semi-Supervised Learning

Towards further incorporating information from unlabelled time series, the area of semi-supervised learning [12] is also of interest. Semi-supervised learning leverages both labelled and unlabelled data, utlising the underlying structure or manifold of the unlabelled samples to improve model performance and generalisation. It, thus, makes the most of manually-labelled data, rendering it a cost-effective and scalable approach for training models on datasets where annotation is costly or limited. Prior work has highlighted the pivotal role of semi-supervised learning in human activity recognition tasks [109], as well as in healthcare [92] and other real-world data scenarios [110], demonstrating its ability to

improve classification with fewer labelled samples. However, traditional semi-supervised learning methods may not fully capture the domain expertise that medical professionals can provide, limiting their practical utility in healthcare time series applications.

Active semi-supervised learning explicitly addresses this gap [111] by integrating the complementary strengths of active learning and semi-supervised learning [112]. In this case, the most informative samples are selected for manual annotation while the remaining unlabelled data contribute to model training through pseudo-labelling or consistency-based regularisation [113]. Such approaches have been shown to improve performance while reducing annotation costs with incremental data acquisition. However, they assume a fixed and pre-configured set of model hyperparameters, which leaves users with the cumbersome task of choosing suitable configurations a-priori. This can hinder the full potential of leveraging unlabelled data, particularly when the learning process itself is dynamic.

Recent studies also highlight robustness challenges in semi-supervised learning, especially under open or non-stationary environments [114]. In healthcare time series, this underscores a key limitation: while both active learning and semi-supervised learning are designed to operate with unlabelled data, they typically assume fixed model configurations and overlook the dynamic interplay between data acquisition, model adaptation, and incremental learning. Consequently, users still face the cumbersome task of pre-selecting model hyperparameters, adding complexity to the already challenging process of data labelling. This emphasises the need for approaches that can systematically explore model configurations alongside incremental data acquisition, while effectively incorporating both human expertise and evolving model insights.

2.5.4 Weak Supervision

Other work using unlabelled data and human annotations has looked at weak supervision to automatically generate training samples from medical data [115]. Work leveraging noisy heuristics and distant supervision [116] has also combined labelling functions and probabilistic label aggregation, which iteratively refines the labelling functions. However, this paradigm typically assumes that labelling functions are sufficiently diverse and only weakly correlated, but this is an assumption that is difficult to guarantee in practice in biosignal settings where expert heuristics often encode overlapping domain knowledge [117]. Moreover, the probabilistic label models rely on accurately estimating the unknown accuracies and dependencies of these functions without access to ground truth, which can lead to miscalibration and error propagation when these assumptions are violated [117]. Additionally, the reliance on weakly supervised labels may introduce substantial noise [117], and recent work on predictive inference under weak supervision highlights that even defining appropriate notions of coverage and validity becomes non-trivial in the

presence of such noisy and partial annotations [118], which can be particularly detrimental in health data.

The trade-offs between the effort required to develop labelling functions for biosignals and the savings in manual labelling are significant, rendering the approach less effective. Specifically, biosignal data often exhibit high inter-subject variability, temporal dependencies, and complex noise characteristics, which make it difficult to encode reliable heuristics as labelling functions without extensive domain expertise and iterative tuning [115]. While leveraging weak supervision can automate label generation, the practical burden of constructing, validating, and maintaining these labelling functions combined with the risk of systematic bias introduced by imperfect heuristics and noisy supervision limits their robustness in clinical and other real-world contexts [117, 116]. Therefore, the constraints around designing labelling functions for biosignals reduce the overall effectiveness of weak supervision.

2.5.5 Summary

In summary, existing work does not fully solve the challenges posed by unlabelled time series data. Addressing this requires not only reducing the burden of manual labelling, but also improving the adaptability and accuracy of ML models to handle the unique temporal characteristics of time series data. By integrating domain expertise more effectively, introducing innovative mechanisms for enhancing data labelling, and promoting automated learning through novel approaches, the potential for advancing ML in real-world applications is substantial. Ultimately, bridging this gap can lead to more efficient, accurate, and effective ML solutions.

2.6 Distribution Shifts in Time Series

Despite the importance and usefulness of enhancing the automation capabilities of the ML model development process, the use and adaptation of pre-trained models remains paramount. This is especially the case in scenarios where a small labelled subset is not sufficient for model training due to the complexities of the underlying data, and so the go-to solution is the fine-tuning of a model pre-trained on similar data.

One such example are ECG datasets, which can be highly indicative of a wide range of cardiac pathologies, including arrhythmias, myocardial infarctions, and other cardiac anomalies [15, 16], and whose labelling process is even more time-consuming than that of other modalities, requiring the expertise of trained cardiologists [119, 120]. Unfortunately, fine-tuning pre-trained models often does not perform as well as it could under out-of-distribution (OOD) conditions, where test data diverge from the training distribution due to differences in populations, in sensors, and in clinical protocols [9, 20] between the

pre-training and deployment datasets.

While transfer learning and domain adaptation methods have been proposed to mitigate such distributional differences [121], they all face common challenges, including modality mismatch, limited labelled data, and inter-subject variability [4, 122]. These methods treat OOD cases under fixed assumptions without considering their granularity or severity. They often apply the same adaptation regardless of differences in shift levels or task complexity, so they often do not generalise to the broad spectrum of OOD data that may occur during fine-tuning [123], as most adopt a coarse-grained binary approach, treating data as either in-domain or out-of-domain. However, in practice, multiple sources of shifts with varying severities frequently overlap [124]. For instance, a model trained on ECG data from young adults might underperform on elderly patients (population shift) or on data from a different device (sensor shift). As such, in this section we review related work on approaches for handling distribution shifts and relevant training paradigms, which have attracted considerable attention.

2.6.1 Transfer Learning

The approach of transfer learning focuses on pre-training a model on one task and fine-tuning it on a related one [17, 125], leveraging learned representations to improve performance on smaller or less representative datasets [18]. This process typically involves reusing the weights of a pre-trained network, namely its learned feature representations or embeddings. These capture hierarchical patterns in the input data and can be transferred across related domains, and they may be used either as fixed feature extractors or further adapted through full or partial fine-tuning, depending on the similarity between the source and target tasks. This is particularly important given prior work demonstrating that models trained directly on small datasets tend to generalise poorly, whereas those trained on larger, more comprehensive datasets typically achieve significantly stronger performance [126]. Additionally, pre-training on large-scale datasets enables models to encode more robust and semantically-meaningful features, improving both convergence speed and sample efficiency in downstream tasks.

The real-world effectiveness of a model is not solely determined by its architecture or training procedure, but it also depends on how extensive, varied, and truly representative the training data samples are in capturing the characteristics of the activities the model will encounter. In this context, transfer learning provides a mechanism to bridge gaps in data availability by transferring knowledge from data-rich domains to data-scarce ones, mitigating the need for costly data collection and annotation [127]. Large-scale studies have shown that better in-distribution accuracy frequently correlates with improved OOD accuracy [128], but these gains remain limited under in-domain scenarios, where the testing and training data share the same distribution.

Moreover, existing methods struggle to consistently handle shifts across different tasks or diverse shift types, thereby constraining their overall generalisability [17]. This can lead to degraded performance when the transferred representations fail to adequately capture target-specific characteristics. In such cases, where the source and target distributions differ substantially, there is also the possibility of negative transfer occurring, This is a phenomenon in which knowledge encoded in the pre-trained model, typically in the form of learned weights or feature embeddings, adversely affects the target task's performance, resulting in lower accuracy and overall effectiveness compared to training a model from scratch [129]. This typically arises when the representations of the source distribution encode spurious correlations or distribution-specific biases that do not generalise to the target distribution [130]. As such, it is imperative to develop strategies that account for distribution shifts while preserving task-relevant information.

2.6.2 Representation Learning

Representation learning aims to improve generalisation by extracting more stable features across changing data scenarios, focusing on capturing robust embeddings [131]. Representation learning transforms raw input data into a structured internal form, often numerical vectors, that capture the most salient and informative patterns and relationships. This automated process reduces reliance on manual feature engineering and allows models to efficiently leverage underlying structures in the data for downstream tasks such as classification or prediction [132].

Prior work has explored learning representations robust to latent and dynamically changing distributions [133] for OOD time series. However, this expects the data to come from a limited set of known conditions, limiting its generalisability to unseen data distributions. Such assumptions are particularly restrictive in real-world settings where distribution shifts may arise from multiple, overlapping sources with varying and unknown severity, making it difficult for models to anticipate all possible variations during training. In parallel, general approaches to ECG representation learning [134] have aimed to learn task-agnostic embeddings transferable across patient populations and diagnostic tasks, but often depend on pretext tasks or handcrafted augmentations that may not fully capture clinically relevant temporal variability in OOD settings. Recent surveys on universal time series representation learning overall highlight that existing methods often rely on fixed training objectives or dataset-specific design choices, which can limit their ability to generalise under complex and evolving temporal shifts [135].

The insights gained from other representation learning domains further contextualise these challenges. In tabular data, existing literature has categorised models into specialised, transferable, and general approaches, reflecting different levels of adaptability across datasets and tasks [136]. Yet, while transferable and general models aim to

learn reusable representations, they often assume relatively stable feature distributions or well-defined adaptation settings. Similarly, multi-view representation learning methods integrate complementary information from multiple sources to learn consistent latent spaces [137]. However, these methods generally rely on aligned or cooperative views and do not explicitly address temporally evolving signals or the presence of overlapping distribution shifts with varying severity, which are central challenges in time series OOD scenarios.

2.6.3 Domain Adaptation

Domain adaptation seeks to transfer knowledge from a source domain to a target domain [19], and prior work has considered several settings depending on data availability and task complexity. For instance, source-free domain adaptation removes the need for access to source data during adaptation [138], while unsupervised approaches address the challenge of transferring models to fully unlabelled target domains. In time series settings, there also exist methods that explicitly model temporal and frequency representations to account for label shifts, enabling transfer under such specific domain discrepancies [139].

Semi-supervised domain adaptation has also been introduced. This operates by assuming access to a small amount of labelled target data, and recent approaches have adapted source data to better match the target distribution, for instance by treating source labels as noisy and refining them from a target-aware perspective [140]. Importantly, the goal of domain adaptation is to bridge the gap between differing distributions so that a model trained on one context can still make accurate predictions on another, even when labelled data in the target domain are limited. Current approaches achieve this through feature alignment using discrepancy-based or adversarial strategies, but they typically assume a uniform adaptation mechanism and do not account for varying shift severity, which is critical in ECG time series.

More recent work has focused on ECG-specific challenges such as class imbalance and patient variability. For instance, methods for class-specific weighted learning enable efficient patient-specific adaptation by leveraging both shared and individual heartbeat patterns [141]. Similarly, imbalanced domain adaptation networks introduce class-sensitive re-weighting and regularisation that is aware of label distributions to improve performance in multi-class settings [142], while work on multi-source adaptation focuses on aligning feature spaces and classifiers across domains without explicitly quantifying the degree of distribution shift [143]. Thus, the emergence of pre-training methods has reduced reliance on large labelled datasets, but performance often degrades under out-of-distribution (OOD) conditions due to overfitting to the source domain [19]. Existing adaptation strategies frequently treat all shifts similarly, leading to suboptimal performance and inefficient use of resources. This highlights the need for methods that explicitly quantify distribu-

tion shift severity and adapt the fine-tuning process accordingly, enabling more robust and efficient deployment across diverse real-world settings.

2.6.4 Hyperparameter Tuning for Adapting to OOD Cases

Automated hyperparameter optimisation offers improved model generalisation across use cases [97], and it has been found to show strong results under domain and subpopulation shifts with small OOD validation sets [144]. As discussed in Section 2.5.1, hyperparameter optimisation determines the set of hyperparameters that yields an optimal model, effectively minimising a predefined loss function on a given dataset, which is a critical step for achieving robust performance [97]. Classical approaches based on Bayesian optimisation in order to model the generalisation performance as a probabilistic function of hyperparameters, enabling efficient exploration of the search space and often achieving expert-level performance with few hyperparameter tuning trials [95, 94]. Recent comprehensive studies also highlight that hyperparameter optimisation remains central to deep learning performance, significantly influencing convergence behaviour and generalisation across tasks [93].

However, its effectiveness for fine-tuning pre-trained models on OOD time series remains underexplored. Hyperparameter selection directly impacts model performance, and while automated optimisation techniques exist, they have different strengths and limitations depending on the task and data [93, 97]. In particular, their potential for adapting models under distribution shifts, especially in time series settings with overlapping and varying severity shifts, has not been fully investigated even though it could improve robustness and overall performance.

2.6.5 Uncertainty Quantification for Distribution Shifts

Uncertainty quantification is imperative towards improving model fine-tuning, especially under distribution shifts. As discussed in Section 2.3, uncertainty estimates provide an effective mechanism for assessing model confidence, distinguishing between epistemic and aleatoric sources of uncertainty, and informing downstream decisions such as data selection and model adaptation and fine-tuning.

An increase in uncertainty between the pre-training and fine-tuning inputs often signals a shift from the original training data distribution, making uncertainty metrics invaluable for guiding model adaptation. Prior work has evaluated the robustness of uncertainty estimates under distribution shifts, showing that many standard approaches degrade in both calibration and reliability when exposed to distributionally-shifted data, while methods that account for model variability tend to perform more consistently [68]. Subsequent work has further highlighted that uncertainty behaviour is not static, but evolves over

time and under varying degrees of shifts, emphasising the need for protocols that reflect such settings [145].

More recent studies have further refined the role of uncertainty in OOD scenarios. Some approaches attribute poor OOD performance in conventional methods to limited diversity in the learned weight space, and propose entropy-based objectives to better capture epistemic uncertainty and improve detection [69]. Other work has demonstrated the importance of uncertainty in high-dimensional and structured medical data, showing that architecture-aware approaches can improve OOD detection performance [70]. Yet, despite these advances, the application of uncertainty-guided fine-tuning in domains such as ECG biosignals [11] remains largely unexplored, particularly in settings involving heterogeneous and overlapping distribution shifts.

2.6.6 Summary

While existing research has made significant progress in improving model adaptation, many methods still fail to account for the severity of the distribution shifts present in real-world time series, particularly in complex domains like ECG signal analysis. By integrating uncertainty quantification into the fine-tuning process, models can become more adaptable and reliable, especially in real-world settings where OOD data are common. This holds the potential to enhance model robustness and performance, ensuring more reliable outcomes in clinical applications where precision is critical.

Overall, existing methods address aspects of distribution shift but fail to account for its severity and heterogeneity, which are inherent to ECG time series. In practice, ECG adaptation must cope with limited labelled data and overlapping shift types, requiring both parameter-efficient updates and adaptive optimisation strategies. This motivates further exploration into how uncertainty can be leveraged as a data-driven signal to quantify shift severity and guide fine-tuning, enabling robust adaptation across diverse ECG OOD scenarios.

2.7 Conclusions

This chapter has reviewed the evolution and significance of time series in biosignal applications, as well as the potential that machine learning techniques hold in enhancing the analysis of such data through uncertainty estimation. It also presented key challenges persisting in the handling of irregular, unlabelled, and out-of-distribution data in real-world settings. These are summarised in Table 2.1.

While significant progress has been made in ML-based time series applications, especially with the advent of wearable sensors and mobile health technologies, research gaps remain in addressing data irregularities and sparsity, particularly for sequence-to-sequence

Table 2.1: Overview of key limitations in time series modelling and the methods proposed in this thesis to address them. The table highlights three central challenges, including sparse observations, limited labelled data, and performance issues in out-of-distribution conditions, and explains how the thesis introduces targeted solutions that improve uncertainty estimation, learning effectiveness, and model robustness.

Challenge	**Learning Approaches**	**Existing Limitations**	**Proposed Solutions in this Thesis**
Sparse time series	Sequence models assuming regular sampling; pointwise uncertainty estimation methods	Inability to properly model sparse observations; degraded sequence level predictions; limited use of uncertainty in sequence to sequence settings	Chapter 3 introduces SQUIREDL, a sequence to sequence framework with uncertainty awareness that accounts for missing data through imputation aware training and improved reliability
Unlabelled time series	Active learning; semi-supervised learning; fixed training and optimisation pipelines	Separation between annotation and model optimisation; inefficient use of labelling effort; limited integration of expert feedback during training	Chapter 4 proposes SALTS, an adaptive framework that unifies data acquisition, automated hyperparameter optimisation, and effective use of unlabelled data in a continuous learning process
OOD data	Model fine-tuning strategies; transfer learning & domain adaptation methods	Sensitivity to distribution shifts; lack of awareness of shift severity; suboptimal adaptation across different data conditions	Chapter 5 develops ADAPTOOD, an uncertainty guided fine tuning approach that quantifies distribution shift severity and enables efficient and robust adaptation across diverse scenarios

tasks. Additionally, existing solutions for unlabelled data often fail to automate key stages of the model development process or effectively incorporate domain expertise from medical professionals, thus limiting their effectiveness. Moreover, fine-tuning pre-trained models for out-of-distribution data offers a significant but still underutilised potential, particularly when uncertainty is leveraged to support adaptation across different settings and populations. Together, these gaps emphasise the need for further automation in the ML pipeline, using uncertainty-guided approaches to build more generalisable, data-efficient, and adaptive algorithms for time series analysis.

Chapter 3

Handling Sparsity in Time Series for Sequence-to-Sequence Estimation

Uncertainty modelling holds significant promise in deep learning [10], particularly in clinical settings where reliable predictions are essential. By quantifying uncertainty alongside predictions, machine learning models can provide additional insight to medical professionals, increasing confidence in automated systems and supporting safer decision-making. Uncertainty estimates are also valuable during model development, offering holistic and interpretable insights that go beyond conventional performance metrics such as accuracy [146].

Despite these benefits, uncertainty estimation in sequence-to-sequence models remains underexplored compared to traditional single-point tasks such as classification or regression [78]. Effective uncertainty modelling in sequences relies on capturing temporal dependencies, which becomes particularly challenging for sparse time series frequently encountered in healthcare [71]. While much prior work has focused on improving predictive accuracy [82], the explicit modelling of uncertainty has received less attention. Existing approaches either focus on single sequences [83] or overlook temporal dynamics [47, 65]. In real-world medical data, models that assume no missing data often struggle in terms of performance, highlighting the need for approaches that can better account for missing observations.

Evidential Deep Learning (EDL) and its extension, Deep Evidential Regression [58], provide a promising framework for uncertainty estimation. Unlike alternatives discussed in Section 2.3, evidential methods estimate evidence directly within a single model, enabling a clear separation of aleatoric and epistemic uncertainty [59]. While computationally efficient, these approaches still face challenges when applied to sparse data in sequence-to-sequence tasks.

Building on this foundation, in this chapter we present uncertainty-aware sequence-to-sequence prediction on sparse healthcare time series data. To do so, we introduce SQUIREDL, an uncertainty estimation framework for sparse sequence modelling via ev-

idential regression [59]. Specifically, to allow it to deliver trustworthy uncertainty estimates, we mitigate for data sparsity by imputing missing values at the input stage using Akima spline imputation [73], and we refine the loss function of the model so that it is aware of imputed vs true data values, treating them with a different weighting. Further, we examine a set of metrics for assessing the success of uncertainty estimations on sequence-to-sequence tasks, offering a comprehensive evaluation framework beyond the use of accuracy alone, especially for regression tasks where uncertainty metrics are rarely examined.

Initially, we validate SQUIREDL and various evaluation metrics using a synthetic dataset, demonstrating the challenges of existing approaches in processing time series with missing data for uncertainty estimation. We further extend it to two datasets for real-world medical applications: hypoglycaemia prediction for type 1 diabetes and COVID-19 admissions.

Regarding our first clinical application, it is important to note that hypoglycaemia occurs when blood sugar levels drop below 70 mg/dl (4mmol/L), i.e. when a diabetic patient takes too much exogenous insulin. Hypoglycaemia presents with confusion, difficulty in concentrating, dizziness and, if severe, loss of consciousness or even a seizure [147]. Treating hypoglycaemia involves consuming sugar to increase levels above 4mmol/L, yet it can occur suddenly and with symptoms often preventing the person from treating their condition. Some patients have access to automated treatments such as the "artificial pancreas" closed-loop system [148], but most diabetics do not have such an access. In these cases, early warning systems with reliable uncertainty estimation can be lifesaving, as they can help diabetics take corrective action promptly. Models for such systems need to have accurate uncertainty estimation, as relying solely on a – potentially wrong – prediction can lead to severe consequences, such as administering unnecessary insulin, which could cause a dangerous hypoglycaemic event. For example, if a diabetic patient receives an alert related to their blood sugar, but the uncertainty estimation is high, they can choose to monitor their levels more closely instead of immediately taking action, potentially avoiding an unnecessary intervention. Continuously incorporating uncertainty estimation can therefore prevent both overreactions and underreactions, ensuring safer and more effective management of their condition. Hyperglycaemia, while not dangerous in the short term, is the key reason why type 1 diabetics have a lifespan of 8-19 years shorter [149] than people who do not have the condition. If average glucose is high over a lifetime, it damages large and small blood vessels, peripheral nerves, the retina of the eye, the kidneys and the heart. Therefore, a long-term benefit of a continuous monitoring system incorporating uncertainty is that it allows patients to maintain lower average glucose levels, knowing that they will be warned if levels fall. This is achieved by providing timely warnings if glucose levels deviate or the warning mechanism shows a high uncertainty. Through our uncertainty-aware metrics, we capture the ground truth hypoglycaemia risk indices with

an absolute error that is 30% lower compared to baselines.

In our second medical application, we predict weekly COVID-19 hospital admissions based on reported infections. COVID-19 trends are widely-reported time series, yet hospitals face challenges in allocating resources under uncertain conditions [150]. Uncertainty estimation can assist in prioritising resources efficiently as a high predicted admission rate with low uncertainty might indicate a significant resource need, while a high rate with a high uncertainty may suggest a moderate resource requirement, subject to change at short notice. Our model is effective in capturing data sparsity, obtaining an uncertainty-aware absolute error 22% lower than baselines, while it is also directly translatable to other infectious diseases with seasonal outbreaks, such as norovirus and influenza [151], to help hospital resource allocations.

Our key contributions for this chapter are summarised below:

- We develop a novel approach based on evidential regression that addresses sparsity in sequence-to-sequence healthcare tasks by incorporating uncertainty estimation at both the input and the model level.
- We present a data imputation strategy that ensures imputed values are effectively handled during model training, improving the quality of uncertainty estimates.
- We introduce new metrics that move beyond accuracy, offering a more comprehensive assessment of uncertainty in sequence-to-sequence tasks.
- We validate our method on both synthetic and real-world data, demonstrating superior uncertainty-aware predictions across different healthcare use cases exhibiting data sparsity.

3.1 Methods

As per Figure 3.1, SQUIREDL consists of three key components: an imputation module, approximating the missing information in the data to a certain extent; an evidential regression module, predicting the distribution of the output; and an sparsity-aware loss function, employing the knowledge of which points are imputed to better account for the ambiguity of these points and assess the uncertainty estimations.

3.1.1 Data Imputation for the Input Series

In real-world medical sensors, incomplete data is common [71], as they suffer from data transmission errors or temporary malfunctions. As such, an intuitive approach would be to interpolate the missing data. However, given the uncertainty this would introduce without reference to the true values, we propose using Akima spline imputation instead [73].

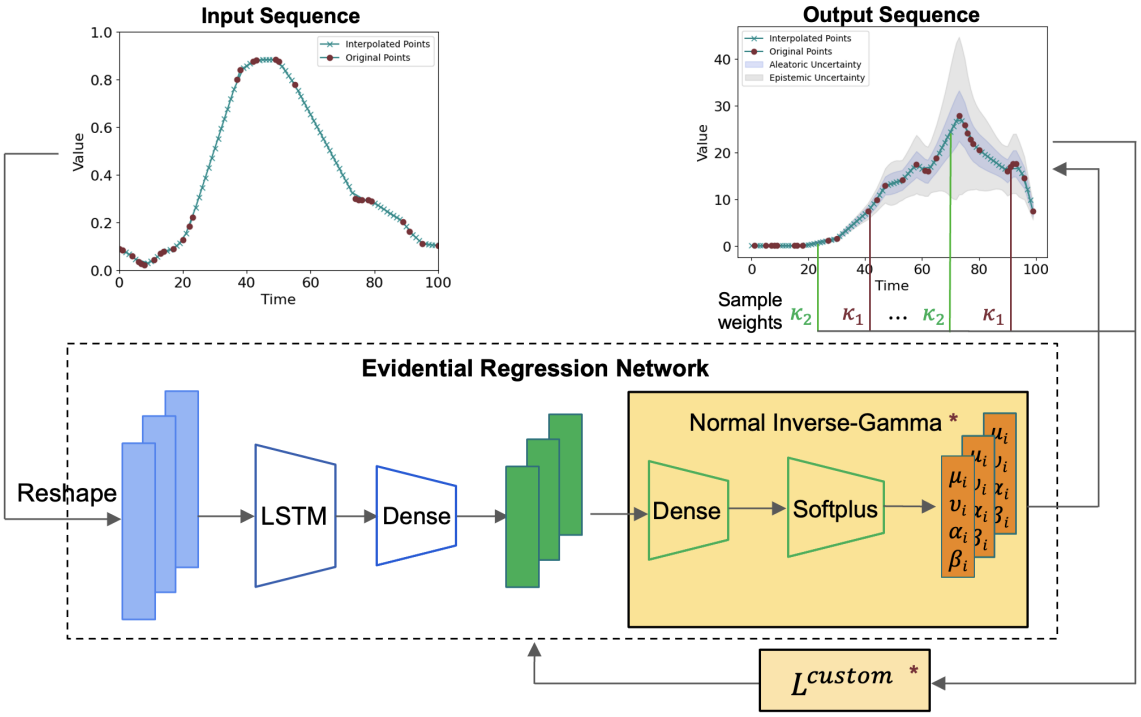


Figure 3.1: Overview of the SQUIREDL model pipeline: an input sequence is imputed using Akima splines, before being reshaped and given to an evidential regression network. This uses an LSTM layer for temporal context capture, a dense layer for feature extraction, a normal inverse-gamma layer for capturing the evidential distribution, and different weights κ (depending on whether the data is observed or imputed) to produce the predicted output sequence enriched with uncertainty insights.

This approach is better at imputing unevenly spaced data compared to traditional linear interpolation, as it has the ability to adapt to the spacing between data points. The imputation is achieved by constructing a piecewise cubic spline $S(x)$ that passes through the given data points, by calculating the slopes $m_i = \frac{y_{i+1}-y_i}{x_{i+1}-x_i}$ for each interval $[x_i, x_{i+1})$ of the sequence [74]. Subsequently, Akima's derivative at x_i is defined as follows:

$$d_i = \frac{|m_{i+1} - m_i|}{|m_{i+1} - m_i| + |m_{i-1} - m_{i-2}|} m_{i-1} + \frac{|m_{i-1} - m_{i-2}|}{|m_{i+1} - m_i| + |m_{i-1} - m_{i-2}|} m_i \tag{3.1}$$

Akima spline imputation places more emphasis on the local variations in the data, which is advantageous in time series segments with rapid changes. Local variations refer to the finer, more immediate fluctuations in the data that occur over smaller time intervals, whereas global variations capture broader trends or patterns over longer periods. Thus, the method's focus on local variations allows it to better capture these changes, especially when the data exhibits abrupt shifts. This is achieved by considering the second derivatives of the curvature of the series [73], making it particularly useful in cases where severe health events occur. For instance, in our CGM case study, where a high hypoglycaemia risk is usually represented with spikes in the data stream, Akima spline

imputation can more accurately reflect these rapid changes (see Section 3.4.2). Thus, it is particularly well-suited for our scenarios, where there is a need to also capture localised and sudden fluctuations.

3.1.2 Evidential regression in SQUIREDL

As displayed in Figure 3.1, SQUIREDL first processes the imputed time series by a Long Short-Term Memory (LSTM) network [64] to obtain the latent representations, which captures long-term dependencies in the time series. The latent representations are further processed through a normal inverse-gamma layer as proposed in [58, 152], which facilitates mapping the embeddings onto the predicted distributions. This layer consists of a dense layer for additional transformation, followed by a Softplus activation function, culminating in the mapping of the output to four parameters $[\mu_i, \upsilon_i, \alpha_i, \beta_i]$. These parameters characterise the distribution corresponding to each time step. The Softplus activation function aims to ensure a valid distribution representation, with positivity for υ, α, and β. The μ_i and υ_i values represent the mean and concentration parameters associated with the normal distribution component, and the α_i and β_i values represent the shape and rate parameters of the inverse-gamma distribution component. Therefore, instead of predicting a single value for each data point, the neural network predicts the parameters of the normal inverse-gamma distribution, providing a probabilistic prediction of the target value that includes both the prediction and its associated uncertainty.

With the predicted probability distribution from the evidential regression, we can estimate both the epistemic (σ_i^e) and aleatoric (σ_i^a) uncertainty of the output, as follows:

$$\sigma_i^e = \frac{\beta_i}{\upsilon_i \times (\alpha_i - 1)} \tag{3.2}$$

$$\sigma_i^a = \frac{\beta_i}{\alpha_i - 1} \tag{3.3}$$

Disentangling the two uncertainties is useful in identifying the uncertainty cause and the possible ways of mitigating it. The epistemic uncertainty is caused due to lack of knowledge and can be reduced by feeding more training data to the model, while the aleatoric uncertainty is due to inherent randomness in the data and cannot be reduced with more data.

3.1.3 Loss Function Used in SQUIREDL Model

The conventional Deep Evidential Regression framework [58] optimises the model using a loss function $\mathcal{L}^{EDL}$, ensuring that the data aligns well with the predicted distribution. It consists of two loss terms: $\mathcal{L}^{NLL}$, capturing the negative logarithm of model evidence, and $\mathcal{L}^{R}$, regularising that evidence and scaling it by a regularisation coefficient λ. Yet, it treats

each data point uniformly and calculates across all points, making it unsuitable for time series with missing data points, common in healthcare applications. To accommodate for this and handle the imputed values within the time series, we propose a custom loss function $\mathcal{L}^{custom}$ which differentially weighs the true and imputed values in the sequences.

For the observed data points, this is represented as:

$$\mathcal{L}^{custom} = \sum_{i=1}^{N} (\kappa_1 \mathcal{L}_i^{NLL} + \lambda \mathcal{L}_i^{R}) \tag{3.4}$$

where i denotes the time index, with a total of N time steps in the sequences, and κ_1 represents the weight given to the observed samples.

Similarly, for the imputed data points, this is represented as:

$$\mathcal{L}^{custom} = \sum_{i=1}^{N} (\kappa_2 \mathcal{L}_i^{NLL} + \lambda \mathcal{L}_i^{R}) \tag{3.5}$$

where i again denotes the time index, with a total of N time steps in the sequences, and κ_2 now represents the weight given to the imputed samples. The weights κ need to be assigned as part of the training process, depending on the applications and the datasets (see Section 3.3.5). Considering the imputed data points, even with different weighting in the loss function, brings notable benefits compared to ignoring them, as it preserves the temporal structure of the time series and maximises data utilisation. This approach allows the model to benefit from additional information provided by imputed data, while also reducing bias in real-world cases where the missingness may not be completely random.

It is worth noting that while SQUIREDL introduces a more sophisticated approach than a generic Deep Evidential Regression model, it does not inherently demand significantly more computational resources for training or deployment. The computational complexity of SQUIREDL, despite incorporating components like our custom loss function $\mathcal{L}^{custom}$, remains on par with that of a standard LSTM-based model. The computational complexity of the additional normal inverse-gamma layer is insignificant as it only consists of a dense layer and an activation function, so the primary source of complexity in SQUIREDL is the LSTM network that, due to its recurrent nature, introduces a computational overhead compared to simpler feed-forward architectures [64]. LSTMs require the maintenance of both hidden states and cell states across time steps, resulting in higher memory usage and more complex calculations. Specifically, at each time step, the LSTM updates both the cell and hidden states, leading to an increased number of operations during the forward and backward passes. This involves additional parameter updates and matrix multiplications, which naturally increases the computational cost relative to non-recurrent models. As such, the computational overhead of LSTMs scales with both sequence length and batch size. SQUIREDL is not significantly more complicated than

that of a generic LSTM model, as the core operations of the network remain unchanged, and the additional steps to compute the distribution parameters and apply $\mathcal{L}^{custom}$ do not substantially impact training or inference times.

3.1.4 Evaluation Metrics

Uncertainty Metrics Overview

Uncertainty quantification is crucial for enhancing the reliability and robustness of any model used in a healthcare setting. While significant progress has been made in uncertainty evaluation for classification tasks, including the introduction of metrics such as the Brier score and the Expected Calibration Error (ECE), these metrics are not directly applicable to regression tasks [153], which is the common case in time series. The Brier score measures the accuracy of probabilistic predictions [154], while ECE quantifies the alignment between predicted probabilities and actual outcomes [155]. With regards to regression, the Paired Euclidean Distances [156] metric is particularly promising, as it offers a straightforward method to quantify the dissimilarity between predicted and actual sequences [156]. A lower value indicates that the predicted and true data points are closer together and thus more similar, however this metric measures prediction accuracy rather than uncertainty. More generally, many commonly used evaluation metrics are designed for specific application domains and primarily assess predictive performance rather than uncertainty. For example, the Bilingual Evaluation Understudy (BLEU) metric [157], widely used in neural machine translation to compare machine-generated translations with human references, illustrates how domain-specific metrics capture task performance despite not intended to quantify uncertainty in numerical health data.

Existing work on uncertainty estimation specific to regression tasks has introduced metrics such as the Negative Log Likelihood (NLL) [158]. This is one of the most commonly-used metrics for regression and is a promising solution for uncertainty quantification, but it reflects uncertainty in an indirect way and it is different from the notion of probability [159]. It measures model quality by quantifying how often the predicted sequence deviates from the true sequence, providing an evaluation of how well the model captures the underlying data distribution [158], but its values are difficult to interpret as they do not have a specific limited range. The Explained Variance Score [160] metric is another promising option for assessing model performance in regression, but it may not fully capture uncertainty in sequences with non-linear relationships and it can oversimplify complex sequence dependencies, overlooking the temporal correlations intrinsic to sequence data. This measures the proportion of variance captured by the model [160], and a higher score indicates that the model explains a large proportion of the variance in the data, suggesting a better fit.

Approaches like Gaussian processes [161] could be useful for sequence-to-sequence

Table 3.1: Formalisation of the proposed uncertainty metrics, examined throughout our evaluation experiments alongside other appropriate existing quantification metrics. These include the interval-based predictive errors (UMAE and UMAPE) computed using both aleatoric and epistemic uncertainty, the ratio of samples whose true values fall outside the predicted uncertainty interval (γ), and the MAE of uncertainty exclusions ($\text{MAE}_{\bar{R}}$).

Metric	**Equation**
UMAE	$\text{UMAE} = \frac{1}{N}\sum_{i=1}^{N} \text{UMAE}_i$; $\text{UMAE}_i = \left\lvert \hat{y}_i \pm \sqrt{(\sigma_i^e)^2 + (\sigma_i^a)^2} - y_i \right\rvert$
UMAPE	$\text{UMAPE} = \frac{1}{N}\sum_{i=1}^{N} \text{UMAPE}_i$; $\text{UMAPE}_i = \frac{\text{UMAE}_i}{\lvert y_i \rvert + \epsilon}$
Ratio of Uncertainty Exclusions	$\gamma = \frac{1}{N}\sum_{i=1}^{N} \gamma_i$; $\gamma_i = \begin{cases} 1 & \text{if } y_i \notin [\hat{y}_i - \sqrt{(\sigma_i^e)^2 + (\sigma_i^a)^2},\ \hat{y}_i + \sqrt{(\sigma_i^e)^2 + (\sigma_i^a)^2}] \\ 0 & \text{otherwise} \end{cases}$
MAE of Uncertainty Exclusions	$\text{MAE}_{\bar{R}} = \frac{1}{J}\sum_{i=1}^{J} \text{MAE}_i$; $\text{MAE}_i = \lvert \hat{y}_i - y_i \rvert; y_i \in \bar{R}$

uncertainty estimation, but pose a challenge in high-dimensional spaces due to computational and modelling complexities. Additionally, they tend to show inferior performance over deterministic deep learning [1], particularly in complex tasks, so they are not an ideal option for uncertainty estimation in such scenarios. Calibration techniques for regression, such as aligning predicted quantiles with observed data distributions [162], have been used for evaluating the reliability of uncertainty estimates at individual points, but they can only assess whether individual predictions are reliable. They do not provide an assessment of the performance across the entire sequence, and assessing uncertainty across entire predicted sequences requires broader evaluation methods that consider aggregate measures and sequence-level characteristics.

Given these constraints of existing approaches, it is thus imperative to assess sequence-to-sequence models for health more robustly, by additionally taking into account new metrics for uncertainty quantification, specific to the needs of sequence-to-sequence tasks.

Uncertainty Metrics in SQUIREDL

Throughout our experiments, we evaluate each approach using both existing and proposed metrics. That's because, as discussed above, most existing metrics for regression tasks aim at scoring the quality of the prediction, not adequately accounting for the requirements specific to sequence-to-sequence tasks. The existing metrics we use include the Negative Log Likelihood (NLL) [158], the Explained Variance Score [160], and the Paired Euclidean Distances [156]. Additionally, we report results using standard metrics, including the Mean Absolute Error (MAE), the Mean Absolute Percentage Error (MAPE), and the

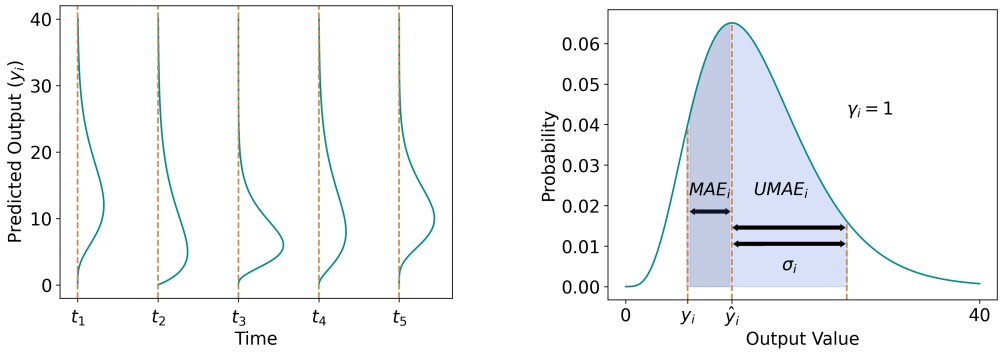


Figure 3.2: Depicting the gamma distributions for a sequence (on the left), showing vertically the probability density function of the predicted output for each time step t_i. Further, depicting visually the UMAE_i and γ_i proposed metrics for one time step i (on the right), where σ_i is the predicted uncertainty given by the standard deviation calculated over the epistemic (σ_i^e) and aleatoric (σ_i^a) uncertainties using $\sigma_i = \sqrt{(\sigma_i^e)^2 + (\sigma_i^a)^2}$.

maximum absolute error [163]. The MAE measures the average magnitude of errors across the data points of the time series without considering their direction, the MAPE expresses the accuracy as a percentage and provides a sense of scale, while the maximum absolute error identifies the largest single deviation between predicted and ground truth values across a sequence.

To enhance the understanding of uncertainty across tasks where it can provide imperative information, we are also examining custom metrics to quantify the success of sequence-to-sequence predictions when factoring in both absolute predictions and uncertainties. As discussed previously, existing uncertainty quantification metrics lack in terms of intuitiveness and completeness. For example, the NLL [158] measures how often the predicted output deviates from the true output, but it is hard to interpret. The Explained Variance Score [160] only measures the variance of the predicted distribution without providing information about the accuracy, as a consistent variance does not guarantee that the predicted distribution is correct. Additionally, the Paired Euclidean Distances [156] metric measures only the distances between the means of the predicted and the true outputs, examining just one aspect of the predicted distribution and lacking in terms of accounting for the variance of the predictions. Therefore, more intuitive measures are necessary to understand how the model performs across the entire dataset, when taking into account both the accuracy and the uncertainty of the predictions. To this purpose, we are proposing the following metrics found below. Of note, the computation of these custom metrics introduces no additional complexity to the solution overall, as they are all based on calculation techniques at the inference/evaluation stage of the system and their calculation takes no more time than the computation of widely used metrics like the Mean Absolute Error (MAE) and others.

A) Uncertainty Mean Absolute Error (UMAE) This metric is similar to the Mean Absolute Error (MAE) seen above, but calculated by adding the uncertainty to the error of the predicted values. Its derivation for each predicted point y_i is done as if the predicted value is at the very edge of the uncertainty threshold. This is seen in UMAE_i of Table 3.1, where $\hat{y}_i$ and y_i are the predicted and true values at time i; σ_i^e and σ_i^a represent the epistemic and aleatoric uncertainty, respectively. The average UMAE can be derived by finding the mean of UMAE_i for all points, where N is the number of points. Of note, this metric will always be larger than MAE, and it quantifies "worst case" predictions in sequence-to-sequence tasks. See Figure 3.2 for a visualisation. It gives a numeric value for the true error in the range predicted for a target sequence that neither the MAE nor the epistemic/aleatoric uncertainty can capture alone.

B) Uncertainty Mean Absolute Percentage Error (UMAPE) This metric is similar to the Mean Absolute Percentage Error (MAPE) existing metric seen previously, but the percentage error is derived using the UMAE as input instead of the original MAE. Similarly to UMAE, this will always be larger than the original MAPE as the error will inevitably be larger when factoring in uncertainty. For each predicted sample y_i the UMAPE_i is calculated as per Table 3.1 and, subsequently, the average UMAPE for a model can be derived for all points. This metric allows understanding what the true "worst case" percentage error is. This is not captured by the normal MAPE. When comparing two solutions that may otherwise look identical, the approach with a lower MAPE may not necessarily be the one with a lower UMAPE too, because if the uncertainty is higher, the prediction may be deemed worse.

C) Ratio of Uncertainty Exclusions The metric indicates what ratio of the true points of an output sequence is outside the thresholds given by the predicted points $\pm$ uncertainty. For instance, if the actual value y_i falls within the predicted range of $\hat{y}_i \pm \sigma_i$, then γ_i will be 0%, and if it lies outside this range it will be 100%. The final γ is assessed by calculating the percentage over all points N in the sequence, as per Table 3.1.

D) MAE of Uncertainty Exclusions The metric estimates how far away the predictions are from the ground truth only for the samples that lie outside the target region $\hat{y}_i \pm \sigma_i$ and returns the MAE for those points. To calculate it, we first find MAE_i for all points y_i falling outside the target area $\bar{R}$ in the output. The metric can then be defined as per Table 3.1, where $\text{MAE}_{\bar{R}}$ indicates the MAE of uncertainty exclusions, and J indicates the total number of points.

3.2 Case Studies

We have validated SQUIREDL using three different datasets, including Lotka-Volterra data, Continuous Glucose Monitoring (CGM) readings for hypoglycaemia detection, and COVID-19 hospitalisations trends. These datasets exhibit varying missing patterns and missing ratios, aiming to validate the effectiveness of our proposed method across varying contexts.

3.2.1 Synthetic Data Exploration

We conduct our initial exploration using synthetic data, which provides the flexibility to control irregularities in time series, facilitating the experimentation with our approach and the proposed metrics. We use Lotka-Volterra data that replicates interactions between two species [164]. Its differential equations generate population dynamics for a prey (i.e., rabbit) and its predator (i.e., fox). When prey populations are high, predator populations increase due to food availability, and vice versa.

3.2.2 Hypoglycaemia Detection with CGM

More than 8 million people have type 1 diabetes, and this is expected to rise to 17.4 million by 2040 [165]. As seen previously, a much-needed application is the real-time prediction of the hypoglycaemia risk based on Continuous Glucose Monitoring (CGM) readings, while incorporating uncertainty. We assess this risk through the Low Blood Glucose Index (LBGI) [166], a well-established measure that is highly sensitive to hypoglycaemia, capturing nuances in glucose variability. This has zero correlation with the opposite range of the blood glucose scale, making it ideal for identifying hypoglycaemia. Traditional LBGI calculation provides retrospective analysis, whereas by employing a sophisticated sequence-to-sequence model with a sliding window (see Section 3.4.2) that learns intricate patterns in the data, we enable timely interventions to prevent hypoglycaemic events. For this case study, we use a dataset from the UVA/PADOVA simulator [167], extracting CGM readings for our input sequence and LBGI indices for the output sequence. This simulator is accepted by the US FDA as a substitute of certain preclinical trials, and has been validated on type 1 diabetics [168] so its data can be considered as real-world for our purpose. Using it, we gather 72 hours of times series for ten patients with readings every 3 minutes. The task we perform is to sequence-to-sequence predict the output (LBGI) and the associated uncertainty, given a sparse input (CGM). Confidence levels are crucial to the output's efficacy, because both low-risk and high-risk predictions with a high uncertainty might indicate previously-unobserved symptoms that require caution. A high-risk prediction with a low uncertainty means that the patient is experiencing a hypoglycaemic event, and only low-risk predictions with a low uncertainty indicate that

the user is safe. Of note, patients may not always choose to consume sugar immediately based on a hypoglycaemic alert, as it depends on the context and their experience of managing their condition, like in times of life including pregnancy when they would like tighter control.

3.2.3 Predicting COVID-19 Hospitalisations

Continuously predicting the number of COVID-19 hospitalisations alongside the relevant uncertainty is another important area, due to the fact that COVID-19 put hospitals under severe pressure [150] and pushed occupancy to record levels. The inferred uncertainty can well acknowledge the resource allocation and help mitigate resource planning issues, which were challenging during COVID-19 surges [150]. To this purpose, we use weekly trends data [169] to predict in a sequence-to-sequence fashion the number of weekly hospitalisations per 100 000 people (output sequence), given the percentage of people testing positive in a population (input sequence) over 130 weeks. Uncertainty estimation is essential in this context, as incomplete or missing data for specific weeks can substantially impact the accuracy and reliability of resource planning. When uncertainty is low, hospitals can prepare the exact number of beds, based on accurate demand prediction. When it is high, hospitals can implement contingency plans for swift resource allocation.

3.3 Experimental Setup

In this section, we discuss the experimental setup used to evaluate our approach. We detail the parameters and conditions under which the experiments were conducted, as well as the baselines, ablation studies, uncertainty estimation methods, and evaluation metrics chosen.

3.3.1 Irregular Time Series for Robustness Testing

We chose to use regular time series and randomly dropped a set of data points to create irregular time series for our experiments. Using regular time series is essential to provide an upper baseline for our proposed method, as the uncertainty estimation does not have a ground truth, especially for the missing segments. Therefore, the model operating on regular time series serves as a reference and provides our upper baseline. To guarantee that the generated irregular time series mimics real irregular time series, we systematically introduce randomness into our dataset (seed 5) by dropping a significant portion of data points. This approach mirrors the inherent sparsity often encountered in practical settings. By randomly discarding indices representing a substantial fraction, as high as 70% of the original dataset, we effectively induce even more irregularity than what is observed in real-world sparsity cases, leading to a mixture of both long and short periods of missing

data. This deliberate introduction of sparsity ensures that our model is robustly tested against the challenges of irregular data.

Of note, SQUIREDL operates by first deducing missing information to some extent, and then applying evidential regression with $\mathcal{L}^{custom}$ to achieve the best possible effect on the missing parts of the data. As such, it is optimised for the missingness type of random missing points, which allows it to recover some of the information through imputation. However, it is also effective in data regions with continuous missing points, though with increased uncertainty in its predictions. This uncertainty arises from the potential presence of outliers and/or severe events in those regions, which the system may have no knowledge of and no prior indication of their occurrence.

We first demonstrate our approach on synthetic data (see Section 3.2.1). We use a simulation [170] to generate 12 500 samples for 10 000 seconds and, for making the series irregular, we drop 4 500 of its 12 500 points. We train for 100 epochs and have a test set of 20%. The test set is regular (no values missing), in contrast to the training set which is irregular for all cases except the upper baseline. Subsequently, to sequence-to-sequence predict the hypoglycaemia risk using CGM readings, we generate the relevant time series [171] as per Section 3.2.2, with a scenario using the Dexcom sensor and adding noise. Computing correlation coefficients, we identify that CGM readings (input sequence) have a correlation with the LBGI risk (output sequence) of -0.478. This is negative because, as glucose decreases, the hypoglycaemia risk increases. The moderate correlation confirms their association, but highlights the task's complexity showing that simple models do not suffice. To make the series irregular, we drop 10 000 of its 14 410 points, making it so sparse that only 30% is kept intact, mimicking the irregularity of challenging scenarios with missing values. We use 30% for test data, 10% for validation, and train for 100 epochs. Last but not least, using the data of Section 3.2.3, we predict weekly hospital admission rates per 100k people (output sequence), given the percentage of people testing positive in a population (input sequence). We use 20% for test data, 20% for validation, and make it irregular for all cases except the upper baseline by randomly dropping the data values for 75 out of the 130 weeks reported.

In addition to the primary experimental configurations described above, we also conduct a supplementary robustness experiment to examine how varying levels of sparsity affect the performance of the proposed framework. Specifically, for the hypoglycaemia prediction task, we use multiple versions of the training data with progressively increasing levels of missingness, ranging from 10% to 90% sparsity. This allowed us to systematically analyse the relationship between data availability and predictive performance and uncertainty. Moderate sparsity levels are handled robustly by the model, while extremely sparse scenarios lead to a gradual increase in prediction error and estimated uncertainty, reflecting the reduced amount of information available for reliable sequence modelling. A detailed quantitative analysis of this experiment is presented in Section 3.4.2.

3.3.2 Baselines for Evaluation Experiments

For our upper baseline, we use a typical EDL model and train it on regular time series (no missing values). For the lower baseline, we train the same model on the irregular/sparse counterpart of the same time series. The upper baseline is a "best-case" scenario, while the lower is a "worst-case". For both, we use evidential regression without imputation and no assignment of different weights to imputed and observed values. The "upper-case" baseline demonstrates the optimal case, so our goal is not to match its performance, but to bring our SQUIREDL's performance as close to it as possible. The "lower case" baseline includes the observed time steps, which helps to capture the authentic data representation without introducing any assumptions or imputations. This mirrors real-world scenarios where unobserved data points are often discarded, and ensures a more accurate reflection of the original sensor readings. The "lower case" setup is trained using irregular time series and tested on regular time series, to better model the practical applications where measurements are irregular and predictions in the future can be regular.

3.3.3 Ablation Studies for Evaluation of Components

To better understand the benefits of the Akima spline-based imputation method in handling irregular time series data, we validate its effectiveness through a detailed ablation experiment. Specifically, we implement a setup using evidential regression without any adjustments to the loss function or differential weighting of imputed and observed points, relying solely on Akima imputation for the missing values. In parallel, we conduct another ablation study to assess the role of the custom loss function $\mathcal{L}^{custom}$, which differentially weighs true and imputed values in the sequences and is a key component of our approach. For this, we implement a setup using evidential regression with $\mathcal{L}^{custom}$ but without any Akima imputation of missing points. Both of these setups are trained using irregular time series data and tested on regular time series data, as described in the baselines of Section 3.3.2.

3.3.4 Comparisons with Other Uncertainty Estimation Methods

To help towards the comprehensiveness of our conclusions, we also perform comparisons with other uncertainty estimation methods, such as MC-dropout, Bayesian techniques, and the approach of learning the variance of predictions as a trainable parameter. While these methods are limited in capturing either the epistemic or aleatoric uncertainty individually, they provide valuable insights for comparison. In the relevant tables of Section 3.4, we report only the relevant uncertainty type (epistemic or aleatoric) for these methods, which helps highlight their performance relative to SQUIREDL in terms of the uncertainty estimates.

The method of MC-dropout [44] works by activating dropout layers during inference to estimate model uncertainty, leveraging the randomness introduced during the prediction process, as discussed in Section 2.3.1. Dropout is applied not only during training but also during inference, causing the network to randomly drop units on each forward pass, resulting in slightly different predictions for the same input. By performing multiple forward passes, the model generates a distribution of predictions that reflects the variability due to uncertainty in the model parameters (epistemic). The mean of these predictions provides the point estimate, while the standard deviation across them quantifies the epistemic uncertainty, indicating the model's confidence in its prediction. This method captures only the epistemic uncertainty, which arises from a lack of knowledge and can be reduced with more data. To implement MC-dropout, we had to make minor adaptations to the system's model: in addition to the LSTM layer described in Section 3.3.5, we added a dropout layer. For a fairer comparison with other uncertainty estimation methods, we also applied Akima spline-based imputation in our studies for MC-dropout and all other methods, expecting enhanced performance over directly applying them to irregular time series.

Bayesian techniques [47] are another helpful way used to estimate the epistemic uncertainty, as discussed in Section 2.3.1. Therefore, we use Bayesian inference to incorporate multiple forward passes of the same input through the model with different weight initialisations. Specifically, we use our LSTM-based neural network, and after training, we generate multiple predictions for the same input. The uncertainty in this case is estimated by computing the standard deviation of these predictions, which reflects the variability in the model's output. This method estimates epistemic uncertainty by considering the distribution of possible outputs from the different weight configurations, which reflects the model's belief about the true output. The standard deviation of the predictions provides an estimate of epistemic uncertainty, once more indicating the model's uncertainty due to lack of knowledge.

On the other hand, learning the variance of the predictions directly as a trainable parameter can only capture the aleatoric uncertainty, which arises from inherent randomness in the data and cannot be reduced with more data. This works by simultaneously predicting both the mean and the variance of the output. The model architecture that we use for this set of experiments remains the same, but the network is trained using an adapted negative log likelihood [158] loss function, which accounts for both the predicted mean and variance. A softplus function is used to ensure that the variance is positive, while the model outputs the mean and the log variance of the predictions. Subsequently, the aleatoric uncertainty is estimated as the square root of the predicted variance. This captures the uncertainty intrinsic in the data itself (aleatoric), rather than the uncertainty in the model (epistemic).

3.3.5 Model Parameters Used for the Experiments

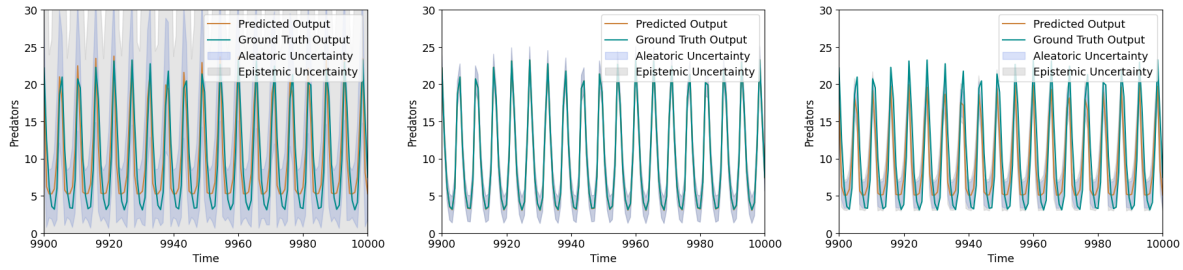
For our SQUIREDL model, we first use an LSTM [64] layer with 64 units to obtain the latent representations. The subsequent dense layer consists of 32 hidden neurons. Following this, the normal inverse-gamma layer is added as per Section 3.1.2. During the model training phase, we use the "Adam" optimiser and minimise the $\mathcal{L}^{custom}$ loss function. The learning rate is 0.001 and the model is trained using an NVIDIA T4 GPU with 16GB of memory and a frequency of 1.59GHz. As per Section 3.1.3, the weights κ of $\mathcal{L}^{custom}$ need to be selected based on the application context, and should be within the range of $[1, 1.5, 2]$. In cases where the output sequences are rapidly changing, the true values are more important to avoid missing any important changes. Therefore, in the hypoglycaemia predictions that are rapidly-varying by their nature, we assign a weight of $\kappa_1 = 1.5$ to the observed values and $\kappa_2 = 1$ to the imputed values, as users need to consider that the ground truth value of the risk index might suddenly change and lie in any part of the region of the predicted value $\pm$ uncertainty. Conversely, in COVID-19 hospitalisation predictions that are more slowly-varying, we assign a weight of $\kappa_1 = 1$ and $\kappa_2 = 1.5$, because users should focus on predicted values in absolute terms to plan ahead as the effects are not as immediate. Depending on how smooth or rapidly-varying the curves are expected to be in each case, the weights need to be assigned accordingly. For experimental purposes, in the synthetic Lotka-Volterra data, where the irregularity and the data spikes are more steady and predictable, we assigned $\kappa_1 = 1$ and $\kappa_2 = 2$, which is another valid choice when the imputation is able to recover most of the smoothly-varying information in the missing part of the sequence. Overall, the choice of weights helps the model optimise for both the accuracy and the uncertainty of the predictions, depending on how time-sensitive the data is and how likely it is to contain sudden spikes.

3.3.6 Evaluation Metrics Used for the Experiments

All experiments conducted, including those regarding SQUIREDL and those regarding the ablation studies, the baselines, and the alternative uncertainty estimation methods, are evaluated using accuracy-based metrics (i.e., MAE, MAPE, Max Error, Paired Euclidean Distances), existing uncertainty metrics (i.e., NLL, Explained Variance Score), and our proposed uncertainty-aware metrics (i.e., UMAE, UMAPE, γ, and $\text{MAE}_{\bar{R}}$).

3.4 Evaluation Results

In this section, we present and analyse the results obtained from our experiments. We compare the performance of SQUIREDL against alternatives, discuss key findings, and highlight observations that support our conclusions.



(a) Lower Case Baseline (b) Upper Case Baseline (c) Full SQUIREDL Proposal

Figure 3.3: Predicted vs ground truth Lotka-Volterra output sequence. In the upper case which uses the full data, it is seen that the default EDL model shows confident predictions, but at the lower case that uses the sparse inputs (∼40% sparsity), the very high uncertainty is meaningless, not providing valuable information. For the same level of sparsity, the SQUIREDL proposal achieves significantly improved performance.

3.4.1 Results on Lotka-Volterra Data

Evaluating Uncertainty-Aware Performance

In Table 3.2 we show the Lotka-Volterra time series experiment results for the different cases of the baselines and the other uncertainty estimation methods. The overall prediction accuracy (indicated through the MAE, MAPE, and Max Error metrics) for the lower baseline is significantly inferior to the upper baseline. This suggests that conventional time series modelling approaches generally cannot effectively process time series with missing points. Further, regarding the predicted uncertainty, it is evident that the upper baseline has very low aleatoric and epistemic uncertainty values (both 1.759) and an excellent Explained Variance Score of 1, while the lower baseline suffers from higher uncertainties (6.167 and 71.425, respectively) and a lower Explained Variance Score of 0.743, suggesting that standard EDL struggles to achieve reliable uncertainty estimation as the extremely large uncertainty values do not provide any useful information. This is possibly due to the model being incapable of capturing reliable temporal dynamics in sparse time series. Similarly, a lower (i.e., better) NLL is achieved in the upper than the lower case (-33.060 vs -7.650), clearly showing the standard model's inferiority in situations of data sparsity.

Generally, SQUIREDL outperforms the MC-dropout, the learned variance, and the Bayesian uncertainty estimation methods, as it achieves better results across multiple metrics. Specifically, it demonstrates an improvement in MAE (3.114 vs 3.652 for MC-dropout, 3.673 for learned variance, and 3.656 for Bayesian), indicating that it provides more accurate point predictions, guided by its custom imputation-aware loss function that leads to smoother and more accurate utility of sparse data information. In terms of relative error, SQUIREDL also achieves the lowest MAPE (0.420), outperforming the other uncertainty estimation methods (MC-dropout: 0.556, learned variance: 0.552, Bayesian:

Table 3.2: Evaluation results for Lotka-Volterra experiments.

Metric	Lower Case	Upper Case	MC-Dropout Baseline	Variance Baseline	Bayesian Baseline	SQUIREDL
MAE ↓	2.962	0.047	3.652	3.673	3.656	3.114
MAPE ↓	0.419	0.010	0.556	0.552	0.572	0.420
Max Error ↓	7.110	0.076	6.118	6.288	5.718	6.851
Paired Euclidean Dst ↓	173.4426	2.637	202.836	204.849	201.049	179.916
NLL ↓	-7.650	-33.060	-2.833	-5.292	-2.802	-8.041
Expl. Var. Score ↑	0.743	1.000	0.648	0.641	0.657	0.727
Aleatoric Unc. ↓	6.167	1.759	-	8.903	-	1.765
Epistemic Unc. ↓	71.425	1.759	1.968	-	1.915	2.491
UMAE ↓	71.941	2.444	4.008	8.897	4.022	3.794
UMAPE ↓	7.948	0.388	0.754	1.759	0.765	0.727
γ ↓	0.000	0.000	0.782	0.000	0.814	0.462
$\text{MAE}_{\bar{R}}$ ↓	0.000	0.000	4.389	0.000	4.268	4.713

0.572), while it also exhibits the best NLL value (-8.041), which highlights its superior probabilistic modelling and ability to estimate the distribution of uncertainties effectively. Regarding uncertainty estimation, SQUIREDL demonstrates a significantly lower aleatoric uncertainty compared to the learned variance baseline (1.765 vs 8.903), suggesting that evidential regression can better isolate the sources of uncertainty in the data. Additionally, the higher epistemic uncertainty estimated by SQUIREDL compared to the MC-dropout and Bayesian baselines (2.491 vs 1.968 vs 1.915) indicates its sensitivity to model uncertainty, which could lead to overestimation in certain cases but also provides a more cautious measure in high-risk scenarios, as also indicated by the UMAE and UMAPE. In terms of the UMAE and UMAPE metrics, SQUIREDL performs similarly to the MC-dropout and Bayesian methods but with slightly better results, while it vastly outperforms the learned variance method (UMAE 3.794 vs 8.897), further confirming that its focus on uncertainty is more reliable.

In this case, it becomes evident that traditional metrics such as the MAE and NLL primarily evaluate the predicted mean or the likelihood of the predicted distribution, but they do not explicitly assess whether the magnitude of the predicted uncertainty is meaningful relative to the actual prediction error. For instance, the learned variance baseline achieves reasonable accuracy but produces a very large UMAE (8.897), indicating that the model compensates for prediction errors by inflating its uncertainty estimates. On the contrary, SQUIREDL maintains a substantially lower UMAE (3.794) and UMAPE (0.727), suggesting that its uncertainty estimates remain better aligned with the true deviations between predictions and ground truth.

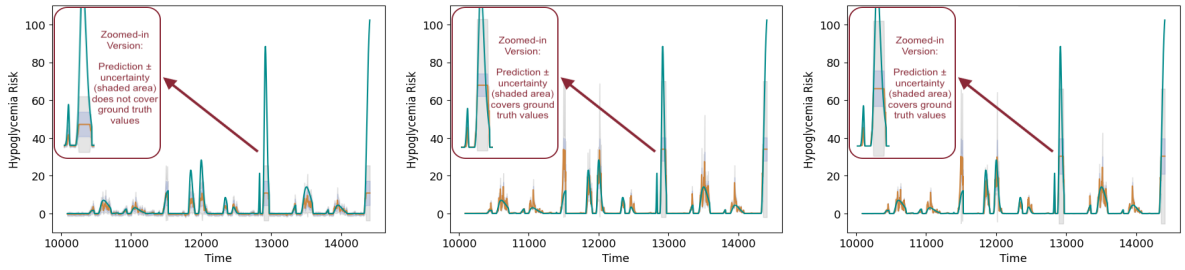
Table 3.3: Ablation study for Lotka-Volterra experiments.

Metric	Lower Case	Without Imputation	Without $\mathcal{L}^{custom}$	SQUIREDL
MAE ↓	2.962	2.791	3.174	3.114
MAPE ↓	0.419	0.348	0.374	0.420
Max Error ↓	7.110	7.966	7.750	6.851
Paired Euclidean Dst ↓	173.526	176.449	193.046	179.916
NLL ↓	-7.650	-7.189	-7.456	-8.041
Expl. Var. Score ↑	0.743	0.745	0.715	0.727
Aleatoric Unc. ↓	6.167	4.118	7.755	1.765
Epistemic Unc. ↓	71.425	36.769	112.152	2.491
UMAE ↓	71.941	36.289	111.202	3.794
UMAPE ↓	7.948	4.976	12.476	0.727
γ ↓	0.000	0.000	0.000	0.462
$\text{MAE}_{\bar{R}}$ ↓	0.000	0.000	0.000	4.713

Component-Wise Analysis via Ablation

According to our ablation experiments seen in Table 3.3, while the accuracy-centric metrics of SQUIREDL show comparable performance to those of the ablations, we observe that SQUIREDL outperforms both the ablation without $\mathcal{L}^{custom}$ and the ablation without the Akima spline imputation in estimating uncertainty. SQUIREDL exhibits smaller aleatoric (1.765 vs 7.755 vs 4.118) and epistemic (2.491 vs 112.152 vs 36.769) uncertainties, as well as a lower Max Error (6.851 vs 7.750 vs 7.966) compared to these cases. It is interesting to note that accuracy-based metrics for SQUIREDL and its ablations show slightly worse performance than the lower case, as for instance the MAE increases (3.114 vs 2.962), but this is due to the emphasis given in optimising for uncertainty. A high accuracy does not always guarantee good model capability, as it can also be very confidently predicting the wrong results, leading to severe consequences such as wrong treatments. Thus, one needs to study the uncertainty too in such applications. When factoring in uncertainty by checking uncertainty-based metrics like the UMAE, UMAPE and NLL, the results of SQUIREDL suggest more confident predictions. In summary, the full SQUIREDL system shows the effectiveness of our design, which combines both imputation and customised loss for effective uncertainty estimation.

Of note, both versions of the ablation study and the lower case show a similar NLL (-7.456 vs -7.189 vs -7.650) and a similar Explained Variance Score (0.715 vs 0.745 vs 0.743), making the comparison challenging with these metrics. As such, we find that solely assessing the uncertainty via the traditional metrics of the NLL or the Explained Variance Score is insufficient and non-interpretable. On the contrary, with the uncertainty-aware metrics of the UMAE and UMAPE, the difference becomes apparent. Comparing the full SQUIREDL model to the version without $\mathcal{L}^{custom}$ or to the version with the Akima



(a) Lower Case Baseline (b) Upper Case Baseline (c) Full SQUIREDL Proposal

Figure 3.4: Predicted vs ground truth hypoglycaemia risk sequence. In the presence of sparse data (70% sparsity), the default EDL model (lower case baseline) correctly predicts the hypoglycaemia risk for low-risk indices, but underestimates it for peak scenarios. Yet, it is at those peaks that the patient is at a real risk, so uncertainty-aware approaches should ensure that these peak predictions are encompassed in most cases within the designated uncertainty region.

spline-based imputation, we notice that the lower UMAE (3.794 vs 111.202 vs 36.289) and UMAPE (0.727 vs 12.476 vs 4.976) indicate less expansive predicted distributions with a lower uncertainty, suggesting the closer proximity of the predictions to the ground truth. Interestingly, the γ of 0.462 and the $\text{MAE}_{\bar{R}}$ of 4.713 are higher than the zero values observed in the ablation cases. This indicates that the proposed model has a greater number of outliers falling outside the exact threshold, whereas the predictions of the other two models fall within this threshold. However, considering the small UMAE and UMAPE values, it is evident that the limited number of outliers is acceptable and reasonable. These outliers are only marginally outside the prediction range, whereas the other two cases may have predictions that are significantly more dispersed, despite being within the region.

This observation also highlights why the γ and $\text{MAE}_{\bar{R}}$ metrics provide additional insight beyond conventional probabilistic measures such as NLL. In the ablation cases, the zero value of γ may initially appear favourable, as all predictions fall within the uncertainty region. However, when considered together with the extremely large UMAE values reaching 111.202, it becomes evident that these models generate wide uncertainty intervals that trivially contain the predictions. SQUIREDL produces tighter uncertainty estimates (UMAE 3.794) while only slightly increasing γ (0.462), thus suggesting a more realistic and informative uncertainty calibration.

Furthermore, as seen in Figure 3.3, the SQUIREDL system demonstrates significant improvements compared to the lower baseline in terms of the predicted uncertainties, with reduced values for both the aleatoric and the epistemic uncertainties. In the upper case, which utilises regular data, the default EDL model exhibits confident predictions. However, in the lower case, where sparse data is used, the extremely high uncertainty renders the predictions meaningless, providing no valuable information. This is also evi-

Table 3.4: Evaluation results for hypoglycaemia prediction experiments.

Metric	Lower Case	Upper Case	MC-Dropout Baseline	Variance Baseline	Bayesian Baseline	SQUIREDL
MAE ↓	2.594	1.896	1.976	2.007	1.877	1.820
MAPE ↓	0.344	0.252	0.580	0.616	0.518	0.192
Max Error ↓	91.538	68.337	60.753	51.328	46.800	69.946
Paired Euclidean Dst ↓	661.512	451.021	346.603	324.856	327.726	460.565
NLL ↓	-7.568	-7.916	-7.928	-5.164	-4.329	-7.873
Expl. Var. Score ↑	0.303	0.658	0.805	0.827	0.819	0.646
Aleatoric Unc. ↓	1.183	0.550	-	8.361	-	0.313
Epistemic Unc. ↓	2.210	2.083	1.005	-	3.018	1.305
UMAE ↓	2.682	2.448	2.776	9.660	4.507	1.900
UMAPE ↓	0.831	0.402	1.247	6.803	2.774	0.249
$\gamma \downarrow$	0.096	0.813	0.350	0.052	0.132	0.860
$\text{MAE}_{\bar{R}} \downarrow$	21.259	1.568	4.650	17.685	9.661	1.737

dent in the versions without $\mathcal{L}^{custom}$ or without the Akima spline-based imputation found in the ablation study, which both highlight the effectiveness of SQUIREDL in estimating uncertainty. It offers a significantly more reasonable uncertainty estimation, resembling very closely the upper case predictions.

3.4.2 Results on Hypoglycaemia Prediction with CGM Data

We further examine SQUIREDL using the CGM hypoglycaemia time series. For this case study, the input sequence consists of the CGM readings and the output sequence consists of the LBGI indices, and we add a sliding window with four time steps to best capture the feature variability inherent in the hypoglycaemia risk application. The choice of a sliding window of four steps aims to guarantee that the intricate patterns in the data are captured by the model, while also enabling a short-term prediction to take place, which is required for rapid decision-making in such a time-critical application.

Evaluating Uncertainty-Aware Performance

Table 3.4 demonstrates that the lower baseline has a higher MAE than the upper baseline (2.594 vs 1.896), as well as a higher MAPE (0.344 vs 0.252), a higher Max Error (91.538 vs 68.337) and a higher value for the Paired Euclidean Distances (661.512 vs 451.021), indicating worse performance for the standard EDL model in the presence of sparse time series. The Explained Variance Score is also lower, and thus worse, in the lower baseline than the upper baseline (0.303 vs 0.658), again suggesting the ineffectiveness of conventional EDL. Moreover, we observe interesting results in SQUIREDL too, with an improved Explained Variance Score (0.646 vs 0.303) compared to the lower baseline, as well as a smaller aleatoric (0.313 vs 1.183) and epistemic (1.305 vs 2.210) un-

certainty. When compared to other uncertainty estimation methods, SQUIREDL shows promising results by offering a balanced approach that excels in minimising uncertainty, as it achieves an epistemic uncertainty of 1.305, which is significantly better than the Bayesian alternative (3.018) and close to that of the MC-dropout alternative (1.005); the slightly lower uncertainty of MC-dropout in this case may be due to its multiple stochastic forward passes, which can enhance calibration in certain cases, but the difference is not substantial in practical terms. The improved UMAE (1.900 for SQUIREDL vs 2.776 for the MC-dropout, 9.660 for the learned variance, and 4.507 for the Bayesian baselines) further reflects SQUIREDL's improved predictive confidence. The lowest UMAPE value of 0.249 reached by SQUIREDL, compared to much higher values for all other methods, also indicates that it produces more consistently reliable predictions.

This comparison illustrates the limitations of accuracy-centric metrics when analysing uncertainty-aware models. Metrics such as MAE and MAPE evaluate only the difference between predicted values and the ground truth, without considering whether the associated uncertainty estimates behave appropriately. Two different models may therefore achieve similar accuracy while producing very different uncertainty distributions. Our proposed UMAE and UMAPE metrics address this by incorporating the uncertainty magnitude into the error evaluation. In this case study, SQUIREDL achieves a lower UMAE (1.900) than the MC-dropout (2.776) and Bayesian alternatives (4.507), indicating that its uncertainty estimates remain more proportional to the observed prediction errors, particularly during peak events that are clinically important. Thus, the new metrics provide information that is key to the uncertainty-aware assessments.

The combination of reduced uncertainties and better explanatory power suggests that SQUIREDL is more effective in estimating uncertainty, showing clearer insights. What is interesting is to also observe the additional information that the proposed evaluation metrics show. The UMAE for our proposal is 30% lower than the lower baseline (1.900 vs 2.682), confirming a better fit. Unlike the other uncertainty estimation methods or the lower case baseline, the value of γ it achieves (0.860) is similar to that of the upper case (0.813), suggesting that its performance is very close to the best case in terms of this metric too. Of note, by manually examining the difference between predictions and the ground truth, we identified that γ is higher for the upper case than the lower case due to the uncertainty for many predictions being zero, whereas the ground truth is marginally different, and thus accounted as outside the uncertainty range.

This example also shows the usefulness of the γ metric when analysing uncertainty calibration. The relatively high γ value of the upper baseline is partly due to several predictions having near-zero uncertainty, meaning that even very small deviations from the ground truth are counted as outside the uncertainty range. Conversely, SQUIREDL produces uncertainty regions that more consistently reflect the variability of the data. Consequently, its γ value of 0.860 remains comparable to that of the upper baseline

Table 3.5: Ablation study for hypoglycaemia prediction experiments.

Metric	Lower Case	Without Imputation	Without $\mathcal{L}^{custom}$	SQUIREDL
MAE ↓	2.594	2.550	1.808	1.820
MAPE ↓	0.344	0.258	0.179	0.192
Max Error ↓	91.538	92.543	68.419	69.946
Paired Euclidean Dst ↓	661.512	672.370	450.760	460.565
NLL ↓	-7.568	-6.064	-5.740	-7.873
Expl. Var. Score ↑	0.303	0.269	0.660	0.646
Aleatoric Unc. ↓	1.183	0.763	0.617	0.313
Epistemic Unc. ↓	2.210	1.334	2.993	1.305
UMAE ↓	2.682	2.425	2.879	1.900
UMAPE ↓	0.831	0.614	0.304	0.249
γ ↓	0.096	0.160	0.499	0.860
$\text{MAE}_{\bar{R}}$ ↓	21.259	13.946	1.297	1.737

(0.813) while still achieving an improved UMAE score (1.900), thus indicating a better balance between prediction accuracy and uncertainty coverage.

Component-Wise Analysis via Ablation

Table 3.5 demonstrates the performance of the full SQUIREDL solution in comparison to versions where certain components have been removed as part of the ablation study. SQUIREDL achieves lower values for both the UMAE (1.900 vs 2.879 vs 2.425) and the UMAPE (0.249 vs 0.304 vs 0.614) metrics, compared to the version without $\mathcal{L}^{custom}$ or the version without Akima imputation. This shows that both modules play a crucial role in improving performance, accounting for both the accuracy and the uncertainty of the output. Of particular interest is that the Akima spline-based imputation methodology has the greatest impact on improving predictions in absolute terms in this hypoglycaemia case study. For instance, the ablation version without the Akima imputation struggles in the accuracy-centric metrics, with a MAE of 2.550, a MAPE of 0.258, and a Max Error of 92.543, indicating that the imputation module is essential and leads to smoother and more accurate imputations that help the model to make better predictions.

Figure 3.4 visualises the performance of SQUIREDL against baselines, highlighting critical scenarios in hypoglycaemia risk prediction. In the presence of sparse data, the standard model (lower case baseline) accurately predicts the hypoglycaemia risk for most indices, but significantly underestimates the risk during peak scenarios. These peaks are critical as they represent real risk periods for patients, and are taken into account in the upper baseline and in SQUIREDL. This demonstrates that SQUIREDL not only matches the upper case performance in low-risk scenarios, but also excels in predicting high-risk peaks. This capability is crucial for real-time alerting systems aimed at helping patients make informed decisions without overloading them with unnecessary alerts. The model's

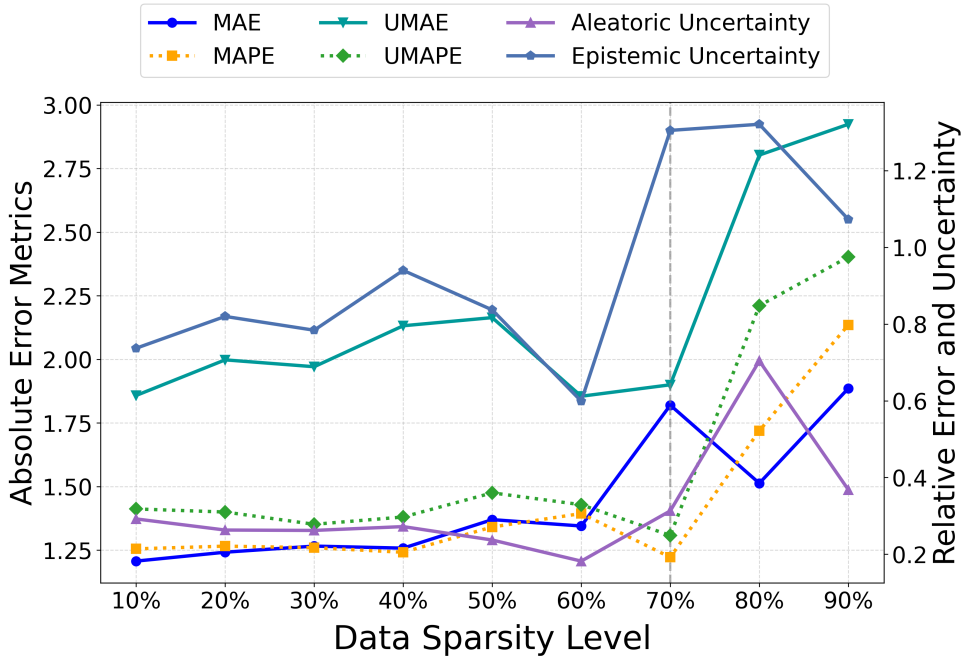


Figure 3.5: Impact of the percentage of data sparsity on model performance. The figure shows the evolution of standard error metrics (MAE and MAPE), uncertainty-aware metrics (UMAE and UMAPE), and aleatoric and epistemic uncertainties as the percentage of missing training samples increases for the hypoglycaemia risk estimation task. Metrics plotted on the left axis correspond to absolute errors (MAE and UMAE), while the right axis shows relative error and uncertainty estimates (MAPE, UMAPE, aleatoric uncertainty, and epistemic uncertainty). The vertical dashed line marks the transition to extreme sparsity conditions ($\geq 70\%$ of missing data), beyond which both prediction errors and uncertainty estimates increase. The model remains robust even at high sparsity levels, while extreme sparsity progressively degrades predictive accuracy and increases model uncertainty.

superior performance in encompassing peak predictions within the designated uncertainty region is a key advancement, ensuring actionable hypoglycaemia risk assessments.

Impact of Data Sparsity on Model Performance

To further examine the robustness of the proposed approach under different levels of irregularity, we conducted an additional experiment on the hypoglycaemia dataset where varying percentages of the training data were randomly removed. Specifically, sparsity levels ranging from 10% to 90% were introduced, allowing us to analyse how progressively missing information affects both predictive accuracy and uncertainty estimation. Figure 3.5 illustrates the behaviour of several key evaluation metrics as a function of the sparsity level.

The results show a gradual degradation in predictive performance as sparsity increases. The MAE remains relatively stable for moderate sparsity levels between 10% and 60%, fluctuating around $1.2-1.4$, but increases noticeably when sparsity exceeds 70%, reaching

approximately 1.89 at 90% sparsity. A similar trend can be observed for the relative error, where the MAPE remains relatively low at values around $0.205 - 0.306$ for sparsity levels up to 60%, but increases sharply at higher levels of missing data, reaching 0.522 at 80% and 0.798 at 90% sparsity. This behaviour indicates that the model is able to maintain stable predictions even when a reasonable portion of the temporal structure is not preserved, and only experiences some difficulty when the amount of available information becomes extremely limited.

The uncertainty estimates provide additional insight into the model's behaviour under sparse conditions. In particular, epistemic uncertainty generally increases with higher sparsity levels, reflecting the model's reduced confidence due to the lack of available observations. The aleatoric uncertainty remains comparatively stable, fluctuating between 0.182 and 0.313 for sparsity levels up to 70%, but increases more noticeably in extremely sparse scenarios, suggesting that the system begins to attribute additional variability to the data itself when the temporal signal becomes fragmented. Correspondingly, the uncertainty-aware metrics UMAE and UMAPE exhibit a gradual increase as sparsity grows. UMAE ranges between 1.855 and 2.164 for sparsity levels up to 60%, but increases to 2.803 and 2.924 at 80% and 90% sparsity, respectively. Likewise, UMAPE remains between 0.278 and 0.360 for moderate sparsity levels before increasing to 0.975 in the most extreme sparsity setting, indicating that extremely sparse time series inevitably limit the amount of information available for reliable sequence modelling. Overall, these observations suggest that the proposed framework remains robust even in conditions with as high as 70% missing data samples, maintaining stable accuracy and uncertainty estimates under high levels of missingness.

3.4.3 Results on COVID-19 Data

Evaluating Uncertainty-Aware Performance

In Table 3.6 we demonstrate our experimental results for the COVID-19 time series case study. The MAE (2.491 vs 1.601), the MAPE (0.419 vs 0.228), the Max Error (4.709 vs 3.790) and the Paired Euclidean Distances (14.513 vs 9.168) metrics are higher in the lower case baseline than the upper case, showing the ineffectiveness of traditional EDL. This may potentially lead to inappropriate resource allocation if used in hospital settings during a pandemic. Additionally, both the aleatoric (1.957 vs 0.481) and the epistemic (2.077 vs 1.844) uncertainties are higher in the lower baseline than the upper baseline, showing that the overall large uncertainty may not reveal trustworthy information, increasing the risks in managing hospital resources using a standard model. This is also reflected through all uncertainty-based metrics too, including the UMAE (5.028 vs 3.013), the UMAPE (0.804 vs 0.456), the γ (0.440 vs 0.400), and the $\text{MAE}_{\bar{R}}$ (3.946 vs 2.556) that also show notable improvements, demonstrating their effectiveness. The fact that uncertainty is

Table 3.6: Evaluation results for COVID-19 prediction experiments.

Metric	Lower Case	Upper Case	MC-Dropout Baseline	Variance Baseline	Bayesian Baseline	SQUIREDL
MAE ↓	2.491	1.601	3.152	2.209	3.241	1.629
MAPE ↓	0.419	0.228	0.415	0.359	0.499	0.232
Max Error ↓	4.709	3.790	5.811	4.452	5.898	3.839
Paired Euclidean Dst ↓	14.513	9.168	17.139	12.354	17.593	9.343
NLL ↓	-6.095	-6.522	-4.850	-5.183	-13.184	-6.389
Expl. Var. Score ↑	0.400	0.622	0.635	0.562	0.638	0.616
Aleatoric Unc. ↓	1.957	0.481	-	8.432	-	0.486
Epistemic Unc. ↓	2.077	1.844	1.149	-	0.848	2.841
UMAE ↓	5.028	3.013	4.231	10.277	4.034	3.944
UMAPE ↓	0.804	0.456	0.638	1.533	0.614	0.591
γ ↓	0.440	0.400	0.840	0.000	0.920	0.080
$\text{MAE}_{\bar{R}}$ ↓	3.946	2.556	3.612	0.000	3.493	3.385

more accurate in the upper baseline, highlights that standard EDL does not handle time series as well when there are missing data points.

These results further demonstrate the need for uncertainty-aware evaluation metrics like those proposed in this work. While MAE and MAPE capture prediction accuracy, they do not indicate whether the predicted uncertainty is well calibrated. For example, the lower baseline shows significantly worse uncertainty-aware metrics, with UMAE increasing from 3.013 to 5.028 and UMAPE from 0.456 to 0.804. This indicates that the predicted uncertainty grows disproportionately relative to the actual prediction error, making the uncertainty estimates less informative on their own even if some traditional metrics appear comparable.

When comparing SQUIREDL to other uncertainty estimation methods and to the baselines, it demonstrates very good performance in the presence of sparse data. Specifically, SQUIREDL achieves a MAE of 1.629 and a MAPE of 0.232, which are very close to the results of the upper baseline, and significantly better than the lower baseline's MAE of 2.491 and MAPE of 0.419. This suggests that SQUIREDL is highly effective at capturing uncertainty and making accurate predictions in sparse data scenarios. In contrast, MC-dropout yields a worse MAE of 3.152 and MAPE of 0.415, which highlight MC-dropout's vulnerability to high variance in uncertainty estimation when the data is irregular. The learned variance baseline produces a slightly improved MAE of 2.209 and MAPE of 0.359, which however is still less effective in managing uncertainty in highly sparse datasets, likely due to its reliance on learning the uncertainty directly by modelling variance during training. The Bayesian baseline performs similarly with a MAE of 3.241 and MAPE of 0.499, further highlighting the challenge of maintaining accuracy and consistency in uncertain environments. While SQUIREDL estimates a higher epistemic uncertainty than the MC-dropout and Bayesian baselines (2.841 vs 1.149 and 0.848), its

Table 3.7: Ablation study for COVID-19 prediction experiments.

Metric	Lower Case	Without Imputation	Without $\mathcal{L}^{custom}$	SQUIREDL
MAE ↓	2.491	2.426	2.220	1.629
MAPE ↓	0.419	0.368	0.327	0.232
Max Error ↓	4.709	4.961	4.760	3.839
Paired Euclidean Dst ↓	14.513	13.411	12.535	9.343
NLL ↓	-6.095	-12.100	-7.635	-6.389
Expl. Var. Score ↑	0.400	0.632	0.605	0.616
Aleatoric Unc. ↓	1.957	0.535	0.492	0.486
Epistemic Unc. ↓	2.077	0.705	2.144	2.841
UMAE ↓	5.028	3.171	4.163	3.944
UMAPE ↓	0.804	0.484	0.622	0.591
γ ↓	0.440	0.880	0.400	0.080
$\text{MAE}_{\bar{R}}$ ↓	3.946	2.702	3.393	3.385

significantly lower aleatoric uncertainty compared to the learned variance baseline (0.486 vs 8.432) suggests that it is more sensitive to model uncertainty and better at distinguishing between noise and true uncertainty, offering a more cautious measure. Notably, SQUIREDL achieves the lowest value of 9.343 for the Paired Euclidean Distances metric, which is very similar to that of the upper baseline (9.168), thus significantly outperforming the alternatives.

Component-Wise Analysis via Ablation

It is worth mentioning that the Explained Variance Scores for the full SQUIREDL proposal and its ablations are very similar (0.616, 0.605, and 0.632), indicating that these versions perform comparably in terms of explained variance. However, when examining the epistemic uncertainty, the ablation without the Akima spline-based imputation methodology shows an improvement (0.705) over the full SQUIREDL model and its variant without $\mathcal{L}^{custom}$, which show similar results (2.841 and 2.144). This suggests that the Akima spline imputation is particularly effective in capturing the variability within the target output sequence, highlighting its importance as a key component of SQUIREDL. Overall, SQUIREDL remains the top performer across several metrics. For example, it achieves a significantly lower UMAE compared to the lower case (3.944 vs 5.028), showing an improvement of 22%. Furthermore, SQUIREDL also demonstrates a reduced UMAPE (0.591 vs 0.804 vs 0.622) and a lower γ (0.080 vs 0.440 vs 0.400) compared to both the lower baseline and the ablation without $\mathcal{L}^{custom}$. These results suggest that it leads to more confident results for sparse sequences, with a greater amount of predictions being within the – much more specific – outputted uncertainty thresholds.

Further to demonstrating the need for uncertainty-aware sequence-to-sequence solutions like SQUIREDL, these results also highlight that conventional metrics alone are

insufficient for evaluating such uncertainty-aware models operating on sparse time series. Models with similar MAE or NLL values may still differ substantially in how informative and well-calibrated their uncertainty estimates are. The proposed metrics (UMAE, UMAPE, γ, and $\text{MAE}_{\bar{R}}$) therefore provide a complementary evaluation perspective, enabling clearer identification of models whose uncertainty estimates remain both realistic and proportional to the observed prediction errors.

3.5 Conclusions

In this chapter we introduced the SQUIREDL system, a novel method for uncertainty-aware sequence-to-sequence predictions in sparse medical time series. Our approach improves deep evidential regression for sparse time series and we demonstrate superiority over baselines, while the approach and the evaluation metrics examined offer a comprehensive view of time series uncertainty in the context of clinical applications. Our methodology opens the door to a new paradigm of applied healthcare studies, which emphasises the importance of uncertainty estimates in sequence-to-sequence predictions.

SQUIREDL provides a more complete picture of potential patient outcomes, rather than solely focusing on model predictions themselves. We demonstrated our approach in two important areas: predicting the hypoglycaemia risk in type 1 diabetes using CGM readings and forecasting hospitalisations from COVID-19 reported case numbers, unlocking a wide range of novel applications of computing in healthcare. Although this chapter has focused on sequence-to-sequence predictions with sparse time series, it does not explore another critical challenge associated with this type of data, particularly that of unlabelled datasets. Thus, further to this contribution, in the next chapter we investigate an approach to automate machine learning model training by maximising the information gained through each human annotation step in an automated way.

Chapter 4

Streamlined Adaptive Learning for Unlabelled Time Series

Nowadays, there are vast amounts of unlabelled time series datasets available, yet the imbalance between data availability and annotation capacity poses a significant challenge, as healthcare ML models require large amounts of labelled data to operate effectively. Sensor datasets require substantial effort to prepare and annotate [90] and this often acts as a bottleneck to the adoption of ML in biosignal applications, as the ML workflow from data annotation to model design and training is not sufficiently abstracted and automated. Labelling raw medical data and integrating them into ML pipelines is a laborious process requiring close expert supervision [2].

Prior efforts in human-in-the-loop annotation, particularly through Active Learning (AL) [90, 91], often assume fixed hyperparameter selections or static tuning, which can hinder performance during data acquisition. Semi-supervised learning is another approach that has been explored to make use of unlabelled data [12], yet it often underutilises domain expertise. While active semi-supervised approaches [92] begin to address this, they still require repeated manual hyperparameter tuning [2]. Automating deep learning with minimal human intervention therefore remains challenging, particularly for healthcare time series due to their temporal and complex annotation requirements [8, 172]. This makes it essential to automate the ML pipeline, focusing on efficiently labelling high-value data while leveraging the inherent structure of unlabelled data.

In this chapter, we introduce SALTS, an end-to-end training framework for classification tasks on healthcare time series that makes the most of a user's annotation budget to build highly-performing models. We first implement an adaptable data acquisition strategy to interactively prompt users to selectively label new data [8] using uncertainty information. At the same time, we delegate hyperparameter tuning [97] to a dedicated refinement step, to automate the key task of choosing the most informative hyperparameters of the learning process. Subsequently, we introduce an automated training phase inspired by semi-supervised learning [12] to enhance predictions by taking into account

the latent manifold on which the data lie through generating pseudo-labels for unlabelled samples, allowing the model to self-finalise and leverage the remaining unlabelled data for more efficient training. SALTS offers greater automation compared to baselines, and only requires initial dataset labelling from users.

Our main contributions for this chapter are summarised as follows:

- We bring hyperparameter refinement concepts to an AL-inspired adaptive data acquisition process, by automatically tuning model hyperparameters iteratively with the data labelling. This requires much less human input pertaining to the model design in an AL setting.
- Our technique makes the most of a pre-set annotation budget, by having progressively more highly-tuned versions of the model being queried for uncertain – and thus informative – data. This reduces the burden on data labellers, while maximising the value extracted from each labelled instance.
- We use unlabelled data through semi-supervised learning after training on as many labelled samples as possible, enhancing performance with no additional labelling.
- We demonstrate SALTS in both binary and multi-class time series classification, using electroencephalography (EEG), electrocardiogram (ECG), and inertial measurement unit (IMU) signals. Compared to alternatives, it brings improvements up to 9% in accuracy in EEG data, up to 6% in ECG data, and up to 22% in IMU data.

4.1 Methods

In this section, we present our proposed methodology for classification tasks. This is illustrated in Figure 4.1 and formalised in Algorithm 1. Our approach features an adaptive training phase operating on an initial small part of the available data by requesting the relevant labels by the user, and an automated part that operates subsequently on unlabelled data with no further user input.

4.1.1 System Overview

Dynamic User Input Phase

SALTS begins with an adaptive data acquisition phase, based on the pervasive assumption in AL literature that there is an initial core set that is already labelled [67, 8]. Once a limited number of samples has been labelled by the user, a preliminary HP selection phase refines the model configuration. The AL loop then smartly selects samples for labelling through uncertainty sampling, prioritising instances with the greatest predictive

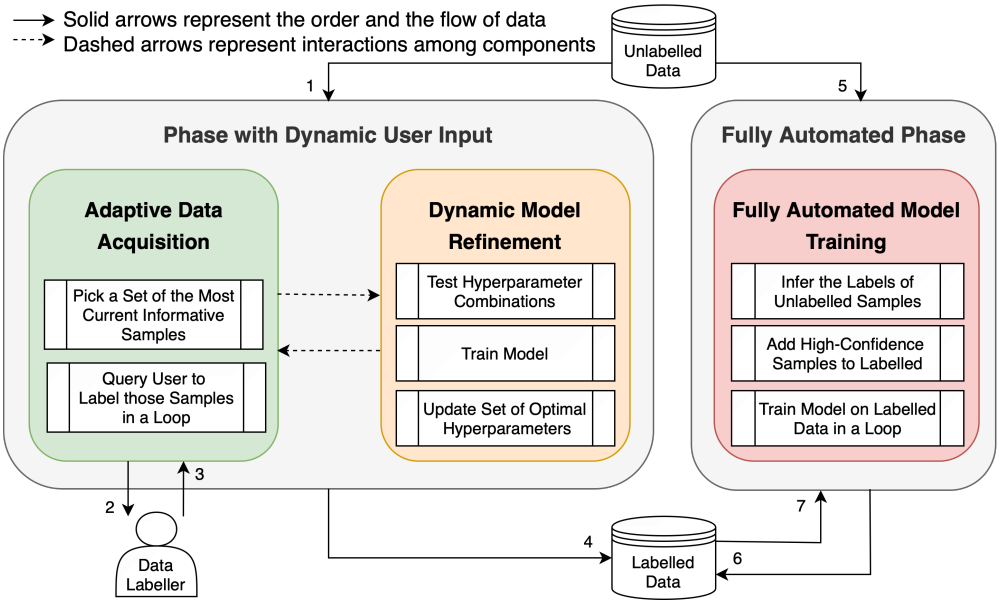


Figure 4.1: Overview of SALTS. A portion of the most uncertain data is labelled at the adaptive data acquisition phase, while the model's hyperparameters are continuously refined. Then, a fully automated phase takes over, where the model pseudo-labels remaining samples, and incorporates them into the training process for classification tasks.

uncertainty and utilising hyperparameters derived from the latest optimisation step. In doing so, the method focuses on samples that are inherently more difficult for the current model to classify, as these are expected to provide the highest information gain. Rather than selecting low-uncertainty instances that are likely similar to already observed data, the approach targets those that challenge the model's decision boundary and can thus improve its generalisation. Importantly, this selection is agnostic to class membership as uncertain samples may originate from any class, and the process naturally emphasises classes or regions of the feature space that are harder to learn. This aims to reduce the required human effort by delegating parts of the process to the automated algorithm. The process continues until the maximum annotation budget is expended, while also running a set of search space exploration trials in between the label acquisition iterations, ensuring that the hyperparameters of the model are continuously tuned.

Automated Training Phase

This phase commences without an explicit initialisation, utilising the data samples that have previously been labelled in the dynamic input phase. Pseudo-labelling is introduced to perform an inference pass on the unlabelled samples and to automatically assign labels to unlabelled samples whose predictions exhibit a low uncertainty [12]. In this sense, uncertainty estimates play a critical role in filtering which samples are deemed reliable enough for pseudo-labelling, while avoiding the propagation of highly ambiguous predictions. The model is then trained on both the pseudo-labelled and the previously-labelled

Input: Unlabelled data D_U
Output: Labelled data D_L, trained model M_L
Parameters: Hyperparameter configuration $\mathcal{H}$, initial hyperparameter iterations limit $\mathcal{T}$, active learning labelling threshold $\mathcal{A}$, per-loop hyperparameter iterations limit $\mathcal{L}$, semi-supervised threshold $\mathcal{S}$
$D_L \leftarrow \emptyset$
$D_{temp} \leftarrow$ small initial part of D_U
$D_{temp} \leftarrow$ request user to label D_{temp}
$D_L \leftarrow D_L \cup D_{temp}$
$\mathcal{H} \leftarrow$ average (midpoint) values in the range of each hyperparameter in $\mathcal{H}$
$M_L \leftarrow$ model.train(D_L, $\mathcal{H}$)
while *hyperparameter trials* $< \mathcal{T}$ **do**
  $\mathcal{H} \leftarrow$ use M_L for exploring the hyperparameter search space to update $\mathcal{H}$
end
while *active learning iteration* $\leq \mathcal{A}$ **do**
  $D_{temp} \leftarrow$ most informative small part of D_U
  $D_{temp} \leftarrow$ request user to label D_{temp}
  $D_L \leftarrow D_L \cup D_{temp}$
  $M_L \leftarrow$ model.train(D_L, $\mathcal{H}$)
  while *hyperparameter trials* $< \mathcal{L}$ **do**
    $\mathcal{H} \leftarrow$ use M_L for exploring the hyperparameter search space to update $\mathcal{H}$
  end
end
while *semi-supervised iteration* $\leq \mathcal{S}$ **do**
  $D_{temp} \leftarrow$ use M_L to infer labels for samples in D_U and pick highly-confident ones
  $D_{temp} \leftarrow$ automatically pseudo-label D_{temp}
  $D_L \leftarrow D_L \cup D_{temp}$
  $M_L \leftarrow$ model.train(D_L, $\mathcal{H}$)
end

Algorithm 1: The proposed SALTS approach is a unified framework that iteratively queries uncertain samples for user labelling, performs hyperparameter tuning, and pseudo-labels high-confidence samples to leverage unlabelled data.

data, facilitating the integration of valuable information from unlabelled samples that an AL solution alone would not be able to capture. The combination of labelled and pseudo-labelled samples leads to the use of all available labels and data structure. This strikes a balance between user involvement and automated processes.

4.1.2 Adaptive Data Labelling and Hyperparameter Tuning

To start the training, our system begins by selecting a small portion of the training data at random and queries the oracle for providing the labels for those samples, inspired by the common step that applies when using AL [67, 8]. The oracle refers to the human-in-the-loop doing the labelling [8], which in our case is performed by an automatic system for testing purposes as we have access to both the labelled and the unlabelled versions of the datasets for our case studies.

Data Labelling

Following on from the initial labelling, the learner in each iteration uses uncertainty sampling [66], enabling progressively more informed decisions on which samples to query the user for labelling next. This strategy prioritises instances for which the model exhibits higher uncertainty, thereby focusing annotation effort where it is expected to yield the greatest improvement in terms of decision boundaries. The uncertainty of each classification using this method is defined as:

$$U(x) = 1 - P(\hat{x} \mid x) \tag{4.1}$$

where x is the instance to be predicted and $\hat{x}$ is the most likely prediction. While we employ uncertainty sampling in this study, the AL acquisition method that can be used by SALTS is flexible and could be adapted to other selection strategies. Similarly, any suitable model can be integrated within our SALTS framework. Once the user annotates a batch of samples, this batch is added to the labelled data pool, and the process repeats in subsequent iterations. A generic AL approach would continue this cycle [8]; however, our framework augments the process by tuning the model hyperparameters, allowing the model to be fine-tuned.

Hyperparameter Space Exploration

To tune the model's hyperparameters, we employ Bayesian optimisation [94], which systematically explores the hyperparameter space by leveraging information from prior trials to guide subsequent ones [93]. Bayesian optimisation maintains a probabilistic surrogate model of the objective function, and uses an acquisition function to trade off exploration of uncertain regions of the space with exploitation of regions believed to yield low error. Unlike alternatives like random search that focus on multiple local optima, Bayesian optimisation aims to identify optimal hyperparameters more reliably by focusing on promising regions of the search space, often resulting in improved model performance consistency [94]. This can reduce the number of evaluations required to reach a near-optimal configuration and provides a structured alternative to less guided search strategies.

The hyperparameters we tune in our solution are strategically diverse. Specifically, the tuner automatically picks the best Conv1D activation function amongst options including eLU [173], ReLU [174], Leaky-ReLU [175] (which facilitates a small gradient for more advanced learning), SeLU (Scaled eLU) [176], and GeLU (Gaussian eLU) [177]. The tuning then extends to the pool size of the MaxPool1D layer, tailoring the network's ability to extract meaningful features. It also tunes the activation function of the dense layer amongst ReLU [174], Sigmoid [178], and TanH, offering flexibility in shaping the network's response to different data characteristics. The optimisation process also selects the most

suitable optimiser from a spectrum including Adam [179], AdaDelta [180], AdaGrad [181], AdaMax [179], and RmsProp [182].

4.1.3 Automated Training on Unlabelled Data

Following the adaptive training phase, our solution stops querying for more labels. At this point, it has learned enough information such that it can confidently classify unlabelled samples. Thus, to train on as many remaining samples as possible, our system proceeds with self-training, a widely used type of semi-supervised learning [12]. This step has no initialisation phase, since a key part of the dataset has already been annotated. Using pseudo-labelling, SALTS picks low-uncertainty predictions and auto-assigns to them the relevant classification labels. For this, we follow the common case of letting the pseudo-labelling run for up to 20 iterations for convergence [183, 14], or until no more low-uncertainty predictions are available (whichever comes first). This lets the algorithm to train on the pseudo-labelled samples and then utilise its knowledge to pseudo-classify further samples [12]. SALTS thus includes both an iterative supervised learning phase and an automated semi-supervised training phase for an automated offering with as little user input as possible.

4.1.4 Focal Loss for Addressing Data Imbalance

We use a focal loss function [184] to address class imbalance in the data, as it is designed to emphasise the harder, misclassified samples at each epoch, potentially also mitigating bias by skewed class distributions [185]. We implement focal loss using the relevant PyPI TensorFlow package [186] to better use the unlabelled data. Focal loss operates by adding a modulating factor $(1-p_t)^\gamma$ to a cross-entropy loss, in order to dynamically scale it and focus on hard-to-classify samples [185]. In our multiclass setting, we use the sparse categorical focal loss, which is defined as follows for a single sample:

$$L(y, \hat{\mathbf{p}}) = -(1-\hat{p}_y)^\gamma \log(\hat{p}_y), \tag{4.2}$$

where $y \in \{0, \ldots, K-1\}$ is the true class label, $\hat{\mathbf{p}} \in [0,1]^K$ is the predicted probability distribution over K classes, $\hat{p}_y$ denotes the predicted probability of the true class, and $\gamma \geq 0$ is the focusing parameter that specifies how much high-confidence predictions contribute to the overall loss [185]. Subsequently, the total loss is computed as the average over all samples. Of note, as γ increases, the easy-to-classify samples are progressively down-weighted, whereas $\gamma = 0$ makes the loss function to behave in the same way as a standard cross-entropy loss. To automate the process of selecting γ, in SALTS we set the HP tuner to optimise this value, allowing the model to focus more rapidly and effectively on informative samples.

4.2 Case Studies

In this section, we present the datasets used in this chapter's evaluation experiments. Each dataset is diverse within the health domain, and classification is performed on a sample-by-sample basis, treating each sample independently. The raw time series signals were utilised as provided, thus incorporating preprocessing steps applied by the original dataset creators, but without any further denoising or feature engineering steps.

4.2.1 EEG Signal Classification

For our first application, we test our proposed method on electroencephalography (EEG) health data. We use the Epileptic Seizure Recognition (ESR) dataset of the UCI ML Repository [187], which is a dataset of EEG recordings modelled as time series. It features 11 500 labelled samples, with each one of them containing 178 attributes. The labels attached to each sample identify if the subject suffered from epileptic seizure or not, further classifying the cases in which the subjects did not suffer in four classes depending on whether they had their eyes open or closed during the data collection process for instance. As the boundaries among the four non-epileptic classes are insignificant, this dataset is used for binary classification, classifying the epileptic seizure class against the rest [187].

4.2.2 ECG Signal Classification

For our second application, we use an electrocardiogram (ECG) heartbeat categorisation dataset [188, 189] from the PhysioNet MIT-BIH Arrhythmia database [190]. This contains 109 446 samples and is characterised by a sampling frequency of 125 Hz, with each signal segmented into 188 points. Its biosignals encompass diverse cardiac events, classified into the classes of normal, supraventricular, ventricular, fusion, and unknown beats.

4.2.3 IMU Signal Classification

For our third application, we evaluate on Inertial Measurement Unit (IMU) health data for human activity recognition. Specifically, we use the accelerometer time series of the Motion Sense dataset [191, 192], which were captured using the accelerometer sensors of a smartphone. This dataset consists of recordings collected from 24 participants who represented diverse demographics, and who were asked to perform 6 activities across 15 trials. While the measurements are associated with specific classes based on the activities performed, it is important to note that they do not involve traditional human annotation. Instead, each participant is assigned a class label based on their activity during the data collection. Despite this distinction, we include this dataset to simulate the AL process on a further scenario. The samples have a sampling rate of 50 Hz, and we use them for

multi-class classification to distinguish activities between six classes: walking, walking downstairs, walking upstairs, jogging, sitting, and standing.

4.3 Experimental Setup

We now discuss the baselines as well as the algorithm settings used for evaluating our proposed SALTS system.

4.3.1 Baselines

To test the efficacy of SALTS we establish some critical baselines: an active learning baseline, a semi-supervised learning baseline, an AutoML baseline, and a hyperparameter tuning baseline. AL strategically selects informative instances for user annotation, expanding the labelled set in an intelligent manner, while semi-supervised learning leverages both user-labelled and unlabelled data to enhance model performance. On the other hand, both AutoML and hyperparameter tuning ensure that the model's configuration is tuned, but they require labelled data to operate. Across all baselines and our system, we run two sets of experiments: one where we label 10% of the samples, and one where we label 20%. These values have been chosen to showcase the alternatives' behaviour at different labelled data percentage allowances, while relevant work like AutoDAL [103] uses the same maximum labelling percentage of 20%. By employing these baselines and providing them all with the same percentage of user-labelled samples, we aim to assess our method's contribution with the same consistent input.

Active Learning Baseline

This baseline uses a standard AL paradigm, and is built using the modAL framework [193], which is a widely-adopted AL solution implemented on top of scikit-learn. This allows us to create a flexible AL workflow to pick the most uncertain samples at each iteration for labelling. For its query strategy, we have chosen an uncertainty-based sampling method [66] that is the same as the one used in SALTS, towards a fair comparison. For the AL implementation, we have set the learner to initially ask the user to annotate an arbitrarily-chosen 1% of the samples in each dataset (seed 111), and then to query them for more labels in a loop until it is fitted to a maximum of 10% or 20% of the dataset (in different experiments). At each iteration, the learner chooses 1% of the samples, corresponding to the most uncertain each time, and the maximum percentage is set so that user does not annotate more samples with AL than with the other baselines. The shortcomings of this baseline are that the user needs to manually set the model hyperparameters to fine-tune them for each task, and that the model is only fitted to as many samples as the user chooses to label (i.e., 10% or 20% in our case). For these AL experiments, we use the

same model hyperparameters as in the semi-supervised learning baseline (Section 4.3.1) for a fair comparison.

Semi-Supervised Learning Baseline

For this baseline, the model is initially given the maximum allowed number of user-supplied labels (either 10% or 20%), and then uses self-training, which is a widely used type of semi-supervised learning [12, 14] that iteratively assigns pseudo-labels to unlabelled samples. The system proceeds with pseudo-labelling the samples whose classification can be inferred with a confidence of > 0.9, choosing this high threshold to ensure that the system only pseudo-labels true positive and true negative samples. Despite leveraging both labelled and unlabelled data, the part of the data which the user is asked to initially label is arbitrarily chosen, so it is rarely the most diverse one. Additionally, users still need to manually set the model hyperparameters. To more strictly put SALTS to the test, we have chosen these hyperparameters carefully for the baseline, unlike how a non-ML expert would choose them in the real world. Specifically, the TensorFlow Keras model used in this semi-supervised baseline starts with a 1D convolutional layer, followed by batch normalisation and max pooling to extract hierarchical features and reduce dimensionality. A convolutional layer is then applied, again followed by max pooling. The subsequent layer is an LSTM layer with 16 units, to capture long-range dependencies. Following this, a dense layer with a ReLU activation is included and the final output layer uses a softmax activation for classification. The model is trained using the Adam optimiser [179] and a focal loss function (see Section 4.1.4), as this combination is the most suitable for integer-encoded labels.

AutoML Baseline

This baseline uses the Auto-SKLearn framework [194], an open-source and widely used AutoML solution built on top of scikit-learn. We choose a randomly-picked (seed 111) 10% or 20% of the data that are manually labelled before feeding them to the system, and then rely on Auto-SKLearn for model development and training. This uses efficient Bayesian optimisation and automatically takes into consideration the performance of previous models on similar types of data for better results. It first initialises its Bayesian optimiser using meta-learning, and then evaluates a set of candidate models on the training dataset. Subsequently, it creates ensembles from these models during its optimisation phase and attempts to identify the factors that will lead to the optimal one. Auto-SKLearn recognises that some models can see a huge impact on their performance depending on the set of hyperparameters chosen and, thus, it features an effective hyperparameter optimisation functionality.

Hyperparameter Tuning Baseline

In this baseline, we use a randomly-picked (seed 111) 10% or 20% of the data for manual labelling by the user, and then invoke Keras Tuner [195] to fit them to a neural network while optimising its hyperparameters. The model starts with a 1D convolutional layer with a dynamically-chosen activation function (eLU, ReLU, Leaky-ReLU, SeLU, or GeLU), and then uses a MaxPool1D layer, where the optimal pool size is automatically determined to enhance feature extraction capabilities. Another 1D convolutional layer follows this, maintaining flexibility in activation functions. Subsequently, batch normalisation and max-pooling layers are incorporated for regularisation and downsampling with a dynamically-chosen pool size, and then another convolutional layer with similar configurations follows. The model ends with a dense layer being added and with its activation function (ReLU, Sigmoid, or TanH) subject to tuning. The optimisation then selects the most appropriate optimiser (Adam, AdaDelta, AdaGrad, AdaMax, or RmsProp) and the most appropriate value for the modulating factor γ of the focal loss function (see Section 4.1.4). This baseline falls behind on the data input level as the labelled instances are arbitrarily chosen, and – similarly to AL – the model is not fitted to more samples than the user labels.

4.3.2 SALTS Algorithm Settings

The model architecture used in SALTS is the same as that of the hyperparameter tuning baseline of Section 4.3.1, and the percentage of the data labelled per iteration is the same as in the AL baseline of Section 4.3.1 for a fair comparison. The further settings we apply are described below.

Hyperparameter Optimisation Trials

For the data labelling phase of SALTS, we use the modAL framework [193]. Ahead of running it, we needed to determine the optimal number of hyperparameter optimisation trials per iteration of the joint data labelling and hyperparameter tuning phase (see Section 4.1.2). We set these to < 3, informed by a comparative study using 20% of the ECG dataset's samples, where the best results were obtained with these values, achieving for instance an average accuracy of 0.969 and an average recall of 0.756. In our analysis, we tested the performance both with fewer (1) and more trials (4) per iteration. The experiment with fewer trials resulted in an accuracy of 0.961 and a recall value of 0.700, while the experiment with more trials achieved an accuracy of 0.966 and a recall of 0.773, indicating that additional trials beyond a certain point yield severely diminishing returns, if any, and just consume more computational resources. Thus, limiting the trials per iteration ensures efficient use of resources without compromising convergence and performance.

Table 4.1: Results from the EEG classification experiments. The values reported are the averages over three repeated trials for each experiment. Both when given a 10% and a 20% labelling budget, SALTS outperforms other methods in all measures.

	Accuracy	**Precision**	**Recall**	**F1 Score**
	10% Labelling Budget			
Active Learning	0.882 ± 0.024	0.813 ± 0.026	0.915 ± 0.015	0.844 ± 0.027
Semi-Supervised	0.965 ± 0.001	0.971 ± 0.003	0.920 ± 0.006	0.943 ± 0.003
AutoML	0.960 ± 0.002	0.953 ± 0.005	0.919 ± 0.011	0.934 ± 0.003
HP Tuning	0.962 ± 0.006	0.963 ± 0.011	0.918 ± 0.019	0.938 ± 0.010
SALTS	**0.978 ± 0.003**	**0.976 ± 0.007**	**0.955 ± 0.011**	**0.965 ± 0.005**
	20% Labelling Budget			
Active Learning	0.893 ± 0.014	0.825 ± 0.015	0.921 ± 0.008	0.856 ± 0.015
Semi-Supervised	0.976 ± 0.001	0.978 ± 0.002	0.945 ± 0.004	0.961 ± 0.002
AutoML	0.965 ± 0.003	0.956 ± 0.001	0.933 ± 0.007	0.944 ± 0.004
HP Tuning	0.969 ± 0.007	0.975 ± 0.003	0.927 ± 0.021	0.948 ± 0.013
SALTS	**0.983 ± 0.003**	**0.984 ± 0.003**	**0.962 ± 0.009**	**0.972 ± 0.006**

Semi-Supervised Settings

For the automated training phase of SALTS that is inspired by semi-supervised learning, the algorithm iteratively pseudo-labels high-confidence – and thus low-uncertainty – samples using the same approach as the semi-supervised baseline seen in Section 4.3.1, towards a fair comparison. In our method, the model hyperparameters are tuned between each data acquisition step, to ensure that the process cooperates more effectively with the data labelling loop. Our approach offers a flexible neural network architecture for classification tasks, combining human and model expertise in an effective way.

4.4 Evaluation Results

We now present the results of our evaluation. All metrics for our experiments have been calculated on the test set and the results can be found in Tables 4.1, 4.2, and 4.3. We run three repeated trials for each experiment and report their average values for greater robustness and consistency.

4.4.1 EEG Classification

In EEG data, SALTS outperforms alternatives in all settings. In test accuracy, it reaches 0.978 when annotating 10% of the EEG samples, outperforming the semi-supervised (0.965), AutoML (0.960) and hyperparameter tuning (0.962) baselines, and achieving a nearly 10% increase in accuracy compared to the AL baseline (0.882). This trend con-

Table 4.2: Results from the ECG classification experiments. The values reported for SALTS and the baselines are averaged over three repeated trials. The values reported for AutoDAL have been visually approximated from Figure 6 of the relevant study [103], depicting the accuracy per percentage of labelled data.

	Accuracy	**Precision**	**Recall**	**F1 Score**
	10% Labelling Budget			
AutoDAL [103]	≈0.93	-	-	-
Active Learning	0.918 ± 0.009	0.559 ± 0.063	0.515 ± 0.016	0.516 ± 0.029
Semi-Supervised	0.930 ± 0.015	0.726 ± 0.030	0.508 ± 0.039	0.543 ± 0.040
AutoML	0.960 ± 0.002	**0.915 ± 0.011**	0.723 ± 0.022	**0.792 ± 0.018**
HP Tuning	0.903 ± 0.027	0.632 ± 0.117	0.473 ± 0.109	0.491 ± 0.078
SALTS	**0.962 ± 0.003**	0.885 ± 0.025	**0.732 ± 0.020**	0.770 ± 0.026
	20% Labelling Budget			
AutoDAL [103]	≈0.96	-	-	-
Active Learning	0.916 ± 0.020	0.553 ± 0.068	0.482 ± 0.067	0.495 ± 0.062
Semi-Supervised	0.950 ± 0.004	0.798 ± 0.093	0.586 ± 0.004	0.599 ± 0.018
AutoML	0.966 ± 0.000	0.924 ± 0.001	**0.764 ± 0.001**	**0.827 ± 0.001**
HP Tuning	0.903 ± 0.081	0.739 ± 0.057	0.629 ± 0.079	0.628 ± 0.135
SALTS	**0.969 ± 0.001**	**0.925 ± 0.021**	0.756 ± 0.007	0.796 ± 0.009

tinues to hold when annotating 20% of the EEG samples, as SALTS reaches an accuracy of 0.983, surpassing the semi-supervised (0.976), AutoML (0.965) and hyperparameter tuning (0.969) baselines, and achieving a 9% increase in accuracy compared to the AL baseline (0.893).

SALTS outperforms baselines in other widely used metrics too, like the recall. With a 10% labelling budget, it reaches recall values beyond 0.95, while baselines range from 0.915 to 0.920. This indicates that it predicts correctly most of the relevant results, in contrast to the baselines. This trend in the recall continues when experimenting with a 20% labelling budget too: SALTS reaches the best value of 0.962, outmatching all baselines. This demonstrates that SALTS uncovers information for EEG samples that each of the alternatives would not do alone given the same human effort.

4.4.2 ECG Classification

In ECG classification, SALTS reaches an accuracy of 0.962 with a 10% user-labelling limit and an accuracy of 0.969 with a 20% limit. For the 10% threshold, this shows that SALTS achieves at least a 6% increase in accuracy compared to the HP tuning baseline, and a 4% and 3% increase compared to the AL and semi-supervised baselines, respectively. Similarly, for the 20% threshold, SALTS achieves a 6% accuracy increase compared to the HP tuning baseline, a 5% increase compared to the AL baseline, and a 2% increase compared to the semi-supervised baseline. Of note, SALTS and the AutoML baseline

Table 4.3: Results from the IMU classification experiments over three repeated trials for each experiment. Both when given a 10% and a 20% labelling budget, SALTS can be seen to outperform other methods in all measures.

	Accuracy	**Precision**	**Recall**	**F1 Score**
	10% Labelling Budget			
Active Learning	0.690 ± 0.045	0.507 ± 0.025	0.532 ± 0.041	0.507 ± 0.035
Semi-Supervised	0.865 ± 0.012	0.814 ± 0.018	0.776 ± 0.012	0.771 ± 0.025
AutoML	0.757 ± 0.013	0.673 ± 0.048	0.626 ± 0.014	0.629 ± 0.022
HP Tuning	0.799 ± 0.028	0.685 ± 0.164	0.685 ± 0.068	0.654 ± 0.091
SALTS	**0.911 ± 0.017**	**0.884 ± 0.030**	**0.903 ± 0.016**	**0.884 ± 0.023**
	20% Labelling Budget			
Active Learning	0.722 ± 0.060	0.534 ± 0.110	0.581 ± 0.071	0.552 ± 0.088
Semi-Supervised	0.874 ± 0.016	0.858 ± 0.023	0.854 ± 0.013	0.842 ± 0.018
AutoML	0.821 ± 0.008	0.729 ± 0.032	0.705 ± 0.007	0.699 ± 0.010
HP Tuning	0.857 ± 0.030	0.801 ± 0.043	0.790 ± 0.062	0.790 ± 0.054
SALTS	**0.925 ± 0.015**	**0.895 ± 0.011**	**0.912 ± 0.013**	**0.898 ± 0.015**

perform similarly in accuracy, indicating that model parameters have the most impact for this data. In terms of the precision, SALTS outperforms all other methods when provided with a 20% labelling budget, reaching a value of 0.925. SALTS is a promising advancement that minimises human input in model development.

In comparing SALTS with AutoDAL [103] (seen in Section 2.5.2), when labelling a total of 10% of the ECG samples, we achieve an accuracy of ≈96%, compared to AutoDAL's ≈93%. This indicates that in settings where annotating unlabelled samples and adjusting model parameters is expensive, and medical experts can only spend fewer hours doing so, SALTS reaches superior accuracy with a single worker machine. When labelling 20% of the samples, we achieve an accuracy of ≈97%, compared to AutoDAL's ≈96%, thus performing better than the state-of-the-art without the computational complexity of a distributed system. This is imperative when compute resources are more scarce, and more effectively reflects the real-world need of medical experts wanting to streamline ML model training, making our solution more scalable and practical for them.

4.4.3 IMU Classification

In human activity recognition with IMU data, SALTS outperforms alternatives in all cases. It reaches an accuracy of 0.911 with a 10% data labelling limit, outperforming the hyperparameter tuning (0.799), AutoML (0.757) and semi-supervised (0.865) baselines, while reaching a significant 22% improvement with respect to the AL baseline (0.690). Similarly, for the 20% labelling threshold it reaches an accuracy of 0.925, thus exceeding the performance of the hyperparameter tuning (0.857), the AutoML (0.821), the semi-

supervised (0.874), and the AL (0.722) baselines.

In terms of the precision, recall and the F1 score, SALTS consistently reaches top values. Yet, the weaker performance of the baselines is also apparent in another, less quantitative area. The AL baseline requires too much effort from the user to tune its hyperparameters and is often less effective when not enough labels are supplied by the labellers, while absence of direct input from medical professionals in semi-supervised models also raises concerns about their suitability for real-world applications. The same applies for the AutoML and the hyperparameter tuning baselines, which cannot function on unlabelled data. This shows that each alternative approach is not sufficient alone, and highlights the importance of SALTS.

4.4.4 SALTS Performance

Across the examined case studies, discussed above and seen in Tables 4.1, 4.2, and 4.3, SALTS performs best with a 20% labelling budget, dynamically choosing the queried data and hyperparameter combinations. For instance, in one of the trials for the ECG experiment with a 20% labelling budget, the system selects Leaky-ReLU amongst the possible activation functions for the 1D convolutional layers (see Sections 4.3.1 and 4.3.2), and 3 as the best pool size for the various MaxPool1D layers. Subsequently, it chooses ReLU as the best activation function for the penultimate dense layer, and continues with the tuning of 1.0 as the most effective γ value for the focal loss function, before finishing with the selection of Adam as the most appropriate optimiser. This dynamic approach, coupled with selecting informative samples, boosts SALTS' performance. The amount of labelled data required for SALTS to consistently outperform baselines is 10%, and our experiments indicate that its performance generally plateaus once 30% of the available samples have been annotated.

4.4.5 SALTS vs. the State-of-the-Art

As per Table 4.2, SALTS outperforms AutoDAL (discussed in Section 2.5) for the same amount of total labelled samples in the ECG data, achieving an increase of ≈3% in accuracy for a labelled data allowance of 10%, and an increase of ≈1% for an allowance of 20%. Similarly, it outperforms the state-of-the-art in AutoML too for most metrics and cases, achieving an increase of ≈2% in accuracy in the EEG dataset, as well as an increase of ≈15% for a labelled data allowance of 10% and an increase of ≈10% for an allowance of 20% in the IMU dataset. This shows that SALTS improves performance, even when the user labels a small part of the relevant data.

4.4.6 SALTS Runtime

SALTS automates model design in the context of AL. Unlike the baselines, which require manual intervention for tasks such as hyperparameter selection, SALTS automates these. Thus, comparing their runtime would overlook the additional human hours that the baselines require. Nevertheless, it is worth noting that SALTS needs ≈3 hours when operating on 20% of the samples of the ECG data. This scales approximately linearly with dataset size and sampling rates, and is based on testing performed on a hosted Jupyter notebook service that uses an NVIDIA GT 710, while having access to 12GB of its RAM. For context, the AL baseline takes ≈0.5 hour, the semi-supervised baseline ≈2.5 hours, the AutoML baseline ≈1 hour, and the HP tuning ≈2 hours. These come with an uncertainty of ±0.5 hour, and can fluctuate depending on the hardware. The execution pace of SALTS allows for finer tuning and reduced human input at the model development.

The automated nature of SALTS ensures consistency, which is evident throughout our experiments, where the variance in the results is minimal. In contrast, both the AL and semi-supervised baselines, which rely heavily on manual input, introduce variability that is less controlled and can add to the burden on data labellers. Therefore, SALTS can handle complex tasks with greater precision while avoiding excessive demands on data labellers. Our experiments demonstrate the efficacy of SALTS, which can uncover information in the health signal datasets during model training that each of the alternatives would not be able to do so alone.

4.5 Conclusions

In this chapter we introduced SALTS, a novel approach for practical learning on unlabelled healthcare time series, incorporating both model and human expertise to minimise user intervention. SALTS is designed as a framework for classification tasks, where uncertainty plays a central role in determining which samples are most informative to label and which can be confidently inferred. Our framework ensures that end users interact with SALTS as little as possible, thus allowing them to make use of the power of ML through its end-to-end automated processes.

SALTS identifies the most uncertain – and thus informative – samples and queries users to label them on the fly, while simultaneously automating a key part of the model development process through hyperparameter tuning. This process is inherently guided by the model's uncertainty estimates, prioritising ambiguous samples for annotation while reserving high-confidence predictions for automatic labelling within the classification setting. SALTS then proceeds to automatically infer the labels for high-confidence samples, letting the model be fitted to a significant amount of additional unlabelled samples with no further user input. We have demonstrated our approach in both binary and multi-class

classification of EEG, ECG, and IMU data, showcasing its applicability to a wide range of healthcare use cases, ensuring domain experts that optimal results will be reached for the samples that they label.

Further to this contribution, in the next chapter we examine that real-world deployment often introduces distribution shifts, particularly in ECG data. Addressing this requires methods that quantify and adapt to such shifts, so we extend our focus from sample-level uncertainty to dataset-level uncertainty under distribution shifts. This is done by introducing a framework that leverages data uncertainty to measure shift severity and guide adaptive fine-tuning of pre-trained models, offering improved performance under diverse conditions.

Chapter 5

Tuning OOD Time Series Models under Distribution Shifts

Data samples collected in one context often differ from those encountered in deployment. Variations in hospitals, sensors, and patient populations can cause distribution shifts that lead models trained on data with even slightly different characteristics to underperform in practice. This problem is particularly important in biomedical time series analysis, where models must operate reliably across heterogeneous clinical environments. ECG signals are crucial in monitoring cardiac conditions and can be indicative of a wide range of cardiac pathologies [15]. With the growing digitisation of healthcare, ML models trained on this type of data have shown promise in detecting arrhythmias, myocardial infarctions, and other cardiac anomalies [16]. Despite this potential, deploying such models in real clinical settings remains challenging, and a key reason for this challenge is the limited availability of high-quality labelled ECG datasets.

Annotating ECG data requires expert knowledge and is time-consuming and expensive [15], making human-in-the-loop solutions less practical for them compared to other modalities [15, 16]. This leads to small datasets that often fail to capture the full variability of clinical settings. While models trained from scratch on these small datasets may perform well within a specific context, they tend to overfit and fail to generalise to new data. Fine-tuning pre-trained models has shown promise in addressing this, as it allows models to leverage general representations learned from large datasets before adaptation.

However, even fine-tuned models can struggle when test data differ from the training distribution. Under out-of-distribution (OOD) conditions, test data diverge from the training distribution due to population heterogeneity, differing sensor types, and variations in clinical protocols [9, 20]. These variations can result in significant distribution shifts, complicating the development of robust and generalisable models for biosignal interpretation [196]. Methods such as transfer learning and domain adaptation aim to mitigate these shifts [121], but they face common challenges: modality mismatch, limited labelled data, and inter-subject variability [4, 122]. Additionally, they often adopt coarse-

grained assumptions that fail to capture the varying severities and overlapping sources of distribution shifts [123, 124].

To address these challenges, in this chapter we propose ADAPTOOD: an adaptation framework that quantifies distribution shift severity, and uses this information to guide model fine-tuning on OOD ECGs. To quantify the severity, we leverage data uncertainty, which reflects how unfamiliar or different a target input is with regards to the pre-training data distribution. The key intuition is that when a model encounters new OOD inputs that are dissimilar to its training distribution, its uncertainty level varies [197]. More specifically, a low uncertainty implies that the OOD dataset lies near the training distribution, indicating a mild shift, while a high uncertainty signals a more severe shift [68, 145]. We fine-tune the pre-trained model based on the guidance of this uncertainty-aware OOD severity quantification mechanism, so that the system adjusts how aggressively it learns from new input. To support this flexible adaptation, ADAPTOOD also incorporates low-rank adaptation and adaptive hyperparameter optimisation. Therefore, this severity-aware approach improves both the effectiveness and the computational efficiency of the adaptation fine-tuning process.

We evaluate ADAPTOOD across real-world ECG applications including arrhythmia detection, cardiac abnormality classification, and age group identification, and we fine-tune the model using datasets that reflect realistic distribution shifts. We compare against transfer learning, supervised learning, and domain adaptation baselines, and also against ablated versions of our approach. Across distribution shifts, ADAPTOOD consistently outperforms alternatives.

Our key contributions for this chapter are summarised below:

- We introduce a novel mechanism that leverages data uncertainty to assess the severity of OOD data shifts in terms of their divergence from the pre-training data.
- We develop an OOD-severity flexible adaptation approach that uses uncertainty information and hyperparameter tuning to achieve better calibration according to the severity of distribution shift, while also incorporating low-rank adaptation to maintain computational efficiency.
- Through comprehensive evaluation in OOD model adaptation, we demonstrate that our method achieves up to 7% higher accuracy and 12.9% higher precision compared to best-performing baselines, as well as efficiency-wise performance gains and consistently strong performance across metrics.

5.1 Methods

In this section, we discuss the formal problem definition of this chapter, provide a system overview of ADAPTOOD, and present the proposed methodology in detail.

5.1.1 Problem Definition

Distribution shifts challenge real-world model deployment [198, 20]. These shifts occur when the joint distribution of inputs and outputs, $P(x, y)$, encountered during testing (D_t), differs from that seen during training (D_s). Common sources include subpopulation shifts [199], where the marginal distribution $P(x)$ changes due to variation across diverse patient groups, or sensor shifts [200], where data is recorded using different devices and leads to variations in $P(x|y)$. Additionally, label shifts arise due to changes in the task-specific label distribution $P(y)$, and temporal shifts due to a distribution drift over time, where $P(x, y)$ evolves due to changes in population health status, sensor calibration, or medical practice [201]. Domain shifts are also important as they can occur when a model pre-trained on one signal, like ECG, is applied to another, such as the photoplethysmogram (PPG) [20]. Finally, contextual shifts driven by changes in recording conditions (e.g., rest, exercise, surgery) alter physiological patterns like the heart rate and blood pressure, challenging generalisation across contexts [202].

5.1.2 System Overview

ADAPTOOD, our proposed uncertainty-guided methodology for OOD model adaptation on ECGs, operates as shown in Figure 5.1. Using an ECG classification model pre-trained on in-distribution data, it focuses on the fine-tuning stage to effectively and efficiently alleviate diverse OOD severity. Specifically, ADAPTOOD includes an uncertainty module to quantify OOD severity levels, a Bayesian hyperparameter tuning module that adapts the model to optimise it to varying tasks, and a low-rank adaptation module (LoRA) to enable parameter-efficient performance. Using the estimated uncertainty to guide the smart unfreezing of model layers, combined with LoRA-based adaptation and Bayesian hyperparameter optimisation, ADAPTOOD fine-tunes the model for robust generalisation on OOD ECGs. In the following subsections, we describe each module.

5.1.3 Uncertainty-Guided Model Adaptation

Uncertainty estimation is fundamental to ADAPTOOD. In our framework, uncertainty reflects the divergence between the pre-training data distribution D_s and the targeted OOD data D_t, by measuring the degree of overlap between them. Specifically, we capture $\sigma(D_t \mid P(D_s))$, representing the uncertainty of the target data with respect to the pre-training source distribution. This measures the degree of overlap (or divergence) between the two distributions, thus serving as a proxy for detecting distributional shifts [145]. A high uncertainty can indicate a severe OOD shift, while a low uncertainty suggests greater similarity to the source distribution [68].

To quantify the shift between D_s and D_t, we start with the data uncertainty estima-

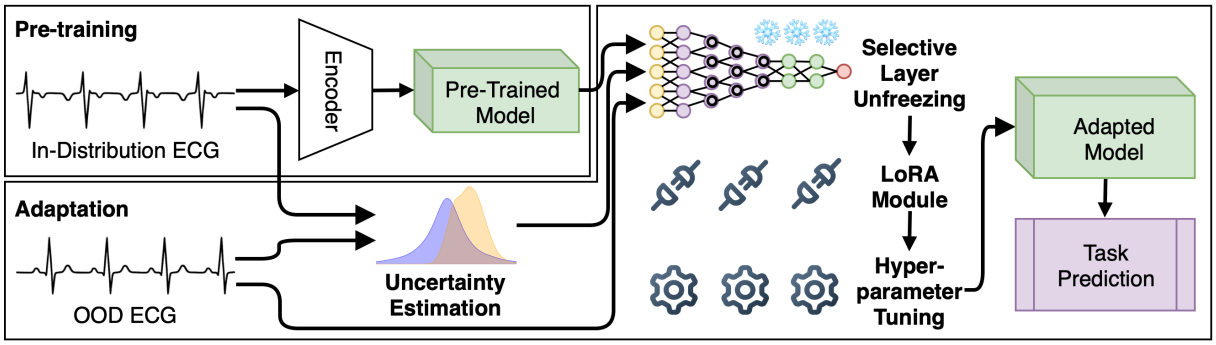


Figure 5.1: Overview of ADAPTOOD. The system adapts a pre-trained model to OOD ECGs by first estimating uncertainty between source and target data, and then using this information to guide selective layer unfreezing. Subsequently, it applies LoRA for efficient fine-tuning and performs hyperparameter tuning for enhanced optimisation across downstream tasks, resulting in a robustly adapted model.

tion. We apply Principal Component Analysis (PCA) to reduce high-dimensional ECG signals into a single dominant component, to improve the effective density estimation while retaining the key variance in the data. We then estimate the uncertainty $\sigma(D_t \mid P(D_s))$ by calculating distribution-level divergence. This is done by leveraging the Mahalanobis distance $(d_m(x))$ and Hellinger distance $(d_h(x))$ as the estimation metrics, discussed below. We apply Gaussian Kernel Density Estimation (KDE) to both the training and OOD dataset to estimate their respective probability density functions, and with these we compute the uncertainty:

$$\sigma(D_t \mid P(D_s)) \leftarrow \{d_h(D_t, D_s), d_m(D_t, D_s)\} \tag{5.1}$$

Mahalanobis distances are a common way of measuring the distance between distributions, considering the correlation between them [203]. To calculate, we compute from the pre-training data the mean vector y and the covariance matrix Σ that represent the centre and spread of the data in the feature space. Using these, we calculate the Mahalanobis distance to measure how far each new batch is from the centre of the known distribution [204]. As such, we use this metric to efficiently quantify OOD severity, where a larger value indicates a higher level of uncertainty, as it reflects the data's deviation from the known distribution and, thus, the OOD severity.

Hellinger distances similarly calculate the similarity or dissimilarity between two probability distributions [205]. They have been previously used to quantify uncertainty in stochastic differential equations with non-Gaussian parameters by minimising the Hellinger distance between empirical probability densities [206]. In our model adaptation case, they can quantify how much an OOD distribution diverges from the training data distribution. Thus, we use them to measure how much the probability distribution of the incoming OOD data deviates from the pre-training data: the greater the distance,

the more severe the OOD shift. For each incoming batch of samples, we compute the Hellinger distance by comparing the probability density of the new samples against the source data.

Choice of Complementary Distance-Based Metrics. We use these two metrics for capturing complementary aspects of uncertainty related to OOD severity. Mahalanobis distances quantify how far OOD batches deviate from the known distribution, incorporating feature correlations and scale [203]. Hellinger distances, by contrast, measure divergence between overall probability distributions, capturing global distributional shifts [205]. Using both gives a more complete picture of OOD severity than a single metric, and we chose them as they account for correlations between variables, unlike simpler metrics like the Euclidean distance, which only focuses on straight-line distances between two points in feature space and does not account for the scale of features [207]. Additionally, unlike alternatives like MC-dropout, which estimates uncertainty by running multiple stochastic forward passes with dropout active [208], these metrics calculate uncertainty before fine-tuning. Thus, they operate pre-inference, allowing for proactive severity estimation.

For computing our final uncertainty metric, given $d_h(D_t, D_s)$ and $d_m(D_t, D_s)$ along with their respective maximum values $d_h^{\max}$ and $d_m^{\max}$, we calculate the normalised distances as follows:

$$\tilde{d}_h = \frac{d_h(D_t, D_s)}{d_h^{\max}}, \quad \tilde{d}_m = \frac{d_m(D_t, D_s)}{d_m^{\max}} \tag{5.2}$$

The combined weighted uncertainty metric is then:

$$\sigma(D_t \mid P(D_s)) = D_{\text{combined}} = w \cdot \tilde{d}_h + (1 - w) \cdot \tilde{d}_m \tag{5.3}$$

Here, $w \in [0, 1]$ is the weight that controls the relative importance assigned to each distance. In our experiments, we set $w = 0.5$ to equally balance the two measures, reflecting a neutral stance when no prior knowledge favours one form of distributional discrepancy over the other. However, this can be adjusted to emphasise either distance metric depending on the characteristics of the source and target domains. Similarly, we standardise both $d_h^{\max}$ and $d_m^{\max}$ to 10, as in the range of datasets we evaluated these values consistently captured the upper bounds of uncertainty without saturation. These aspects remain configurable, useful if applied to datasets with substantially different statistical or uncertainty properties. The combined uncertainty score D_{combined} is then used to inform the degree of selective fine-tuning, as described in the next section.

5.1.4 Selective Layer Unfreezing

Following the OOD severity estimation, as described above, ADAPTOOD uses this information to enable effective representation and parameter transfer from the pre-trained to

Table 5.1: Distribution shifts across datasets in comparison to the PhysioNet CinC pre-training dataset.

	Sensor Shift	**Population Shift**	**Temporal Shift**	**Label Shift**	**Modality Shift**	**Dimensionality Shift**
ECG MIT-BIH	✓	✓	✓			
ECG PTB-DB	✓	✓	✓			
ECG MIMICPERformTT	✓	✓	✓	✓		
PPG MIMICPERformTT	✓	✓	✓	✓	✓	
ECG CODEtest	✓	✓	✓	✓		✓

the fine-tuning encoder. Existing methods typically fine-tune either the entire model or a fixed subset without considering how severe the OOD shift is, leading to unnecessary computational cost and reduced performance due to overfitting when the severity is low or underfitting when it is high. To address this, ADAPTOOD initially freezes all layers of the loaded model and then selectively unfreezes a subset based on the OOD severity. Only the necessary parts of the model get adapted, while preserving the original learned representation. This ensures efficient adaptation without the risks of full retraining, such as excessive cost or catastrophic forgetting of in-distribution knowledge.

The number of layers to be unfrozen is determined dynamically according to the measured uncertainty between the source and target distributions. Intuitively, higher uncertainty indicates a larger distributional shift and therefore requires adapting a greater portion of the network, whereas lower uncertainty suggests that the pre-trained representations remain largely applicable and only minimal fine-tuning is required. To determine the appropriate number of layers to unfreeze, ADAPTOOD combines linear and exponential strategies based on the measured uncertainty, enabling both proportional adaptation and greater sensitivity to severe distributional shifts.

Importantly, the layer unfreezing process proceeds from the output layers toward the input layers. In other words, the deepest layers of the network (those closest to the classifier) remain trainable, while the earlier layers that capture more general features remain frozen. This strategy is motivated by the hierarchical nature of deep neural networks. Earlier convolutional layers tend to learn low-level and general features, such as basic signal patterns, which are typically transferable across related datasets and tasks [1]. In contrast, later layers encode higher-level and more task-specific representations, making them more suitable for adaptation during fine-tuning [135]. By unfreezing layers starting from the top of the network, the model can adjust high-level representations to the new dataset while preserving the stability of previously learned low-level features. As the number of unfrozen layers increases, the fine-tuning process gradually extends toward earlier parts of the network, enabling controlled adaptation of the pre-trained feature hierarchy when larger distribution shifts are encountered.

Linear Calculation. In the linear approach, the number of layers to unfreeze is proportional to the normalised uncertainty between the source and target domains. Given a total number of layers L, this is done as follows:

$$L_{\text{linear}} = (1 - D_{\text{combined}}) \cdot L \tag{5.4}$$

Exponential Calculation. To allow for more sensitivity to uncertainty, an exponential decay function is applied using a decay factor $\alpha > 0$, which governs how rapidly the unfreezing rate drops with increasing uncertainty. The number of layers to unfreeze is:

$$L_{\text{exponential}} = L \cdot e^{-\alpha \cdot D_{\text{combined}}} \tag{5.5}$$

Final Calculation. The final number of model layers that are made trainable by ADAPTOOD is taken as the mathematical mean between the linear and the exponential-based approach:

$$L_{\text{final}} = \frac{L_{\text{linear}} + L_{\text{exponential}}}{2} \tag{5.6}$$

Using this strategy, ADAPTOOD customises the extent of model adaptation according to the OOD severity. By dynamically unfreezing layers starting from the deepest layers closest to the classifier, the framework prioritises adapting higher-level, task-specific representations while preserving earlier layers that capture more general and transferable signal features. Of note, once the trainable layers are selected, the final layers are modified to match the target output shape, ensuring architectural compatibility with the new inputs and outputs.

5.1.5 Low-Rank Adaptation

To make the model adaptation process more effective for OOD scenarios, ADAPTOOD also incorporates Low-Rank Adaptation (LoRA) [209]. We add LoRA support to all the Conv1D layers found in the model, as these constitute the majority of the architecture's core processing units for capturing local temporal patterns. This is especially important in our ECG tasks: since Conv1D layers are responsible for learning feature representations across time steps [210], adapting them directly allows focusing the model's capacity where it matters the most, while keeping the rest of the architecture intact.

With LoRA, we also significantly improve the parameter efficiency of the adapted model. Instead of fine-tuning all parameters, LoRA freezes the pre-trained weights and introduces lightweight trainable components in the form of low-rank matrices [209]. As reflected in the model summary outputs throughout all of our experiments, this results in a noticeable reduction in the trainable parameter counts. Across all datasets, we observe that LoRA consistently leads to smaller model sizes in terms of memory footprint, as per

Section 5.4.4, which is beneficial for deployment and reduces the risk of overfitting when working with limited target data.

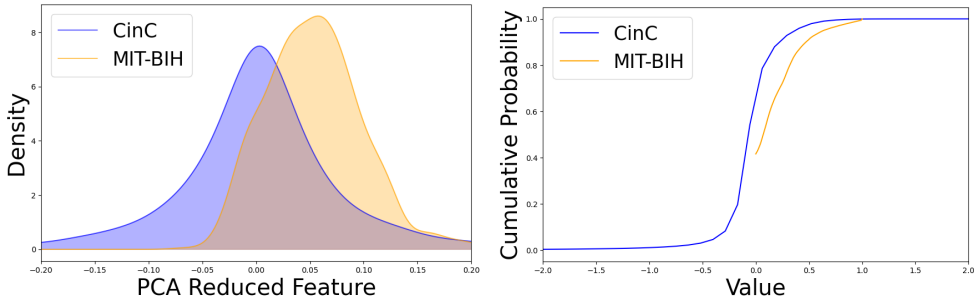
5.1.6 Hyperparameter Optimisation

We further optimise the model through a dynamic hyperparameter tuning strategy, making it even more tailored to the target task with no need for manual adjustments. This is necessary as the hyperparameters performing well during pre-training do not always translate to optimal results under OOD. To this end, we use Bayesian optimisation [94], a widely used approach that iteratively leverages information from prior trials to guide subsequent exploration [93]. Unlike less guided strategies such as random search, which focus on multiple local optima, this targets promising regions of the search space, yielding more consistent performance [94]. This is valuable in OOD settings, where the data distribution differs from that seen during pre-training, as it enables efficient identification of hyperparameter configurations that best accommodate these shifts.

We define a search space over the key dimensions of the optimiser and the learning rate. The optimiser is selected from a pool including Adam [179], AdaDelta [180], AdaGrad [181], AdaMax [179], and RMSprop [182], representing a diverse spectrum. Each has differing behaviours on sparse gradients and noise sensitivity, which can affect adaptation performance in OOD settings. For learning rate selection, we sample from a continuous range using logarithmic sampling to ensure finer-grained search. ADAPTOOD explores the search space for 10 trials, representing a pragmatic trade-off between thoroughness and computational efficiency, given that each trial trains the model for multiple epochs to evaluate varying hyperparameters. In our experiments, this was sufficient to observe meaningful improvements while keeping resource use reasonable. Once the search concludes, we retrieve the best configuration based on validation accuracy for ADAPTOOD's downstream tasks.

5.1.7 Fine-Tuning

After incorporating the above adaptation mechanisms into ADAPTOOD, we proceed to the final stage: fine-tuning the model on the small labelled OOD dataset. During this phase, only the layers deemed trainable by the uncertainty-aware selection process are updated, ensuring focused and efficient adaptation. The model is trained for 20 epochs to balance performance gains with the risk of overfitting. This completes ADAPTOOD's model transfer process.



(a) KDE Plot (b) CDF Plot

Figure 5.2: Distribution shift visualisations between the pre-training data and the MIT-BIH arrhythmia database.

5.2 Case Studies

In this section, we examine the case studies on which we focus our experiments, and we particularly describe the datasets used and the distribution shifts occurring in each case.

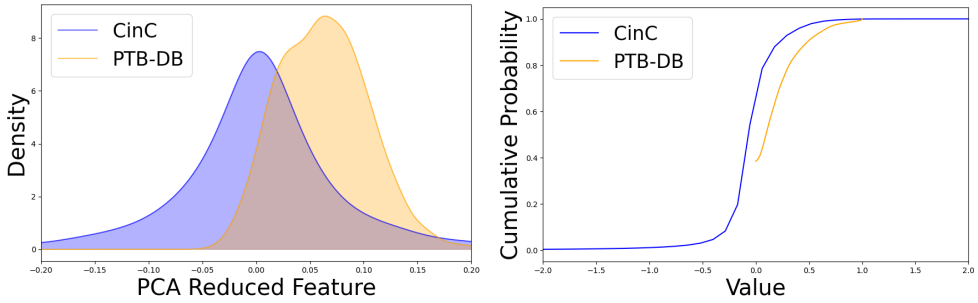
5.2.1 Datasets

We focus on ECGs, as they provide structured temporal signals widely used to diagnose cardiovascular health. Their dynamic and subtle variations make them ideal for evaluating model adaptation. To keep the task realistic and highlight the need for effective adaptation, we fine-tune on only a small subset of each OOD dataset.

PhysioNet Computing in Cardiology (CinC) 2017. This dataset contains 8 528 single-lead ECG recordings [211, 212] and the version used [213] is sampled at 125 Hz. We use it to pre-train a CNN model, which serves as the foundation for subsequent ADAPTOOD experiments. The pre-training task involves classifying the signals into atrial fibrillation and non-atrial fibrillation. As a clinically relevant dataset, this was selected for pre-training due to its well-defined diagnostic labels and task specificity, which provide a strong inductive prior for downstream cardiac classification tasks.

PhysioNet MIT-BIH Arrhythmia Database. This is derived from MIT-BIH's arrhythmia database [190, 214], and contains 109 446 ECGs sampled at 125 Hz and segmented into 188 time steps. We use a single-lead version [189] and randomly select 1 000 samples for fine-tuning to simulate a data-limited setting where training from scratch is infeasible. This tests ADAPTOOD under constrained data, with our classification task distinguishing healthy from non-healthy cases. For evaluation, we use the full test set of 21 892 samples.

PhysioNet PTB-DB Database. This dataset is extracted from the PTB diagnostic database [215, 216] and the version used [189] includes 14 552 single-lead ECGs, sampled at 125 Hz. We retain only 1 000 randomly-selected samples so that the task is realistic



(a) KDE Plot (b) CDF Plot

Figure 5.3: Distribution shift visualisations between the pre-training data and the PTB-DB database.

for model adaptation, and the goal is to distinguish healthy from non-healthy samples.

MIMICPERform ECG Dataset. This dataset is extracted from the MIMIC-III waveform database [217], and consists the MIMICPERform training and the MIMICPERform testing set [218, 219]. The training set contains recordings from 200 patients during routine care, while the testing set includes 200 different patients. The task is to classify samples as adult or neonates, and to better showcase realistic low-resource settings, we further limit the training set to 100 samples, examining adaptation when large-scale training is not feasible.

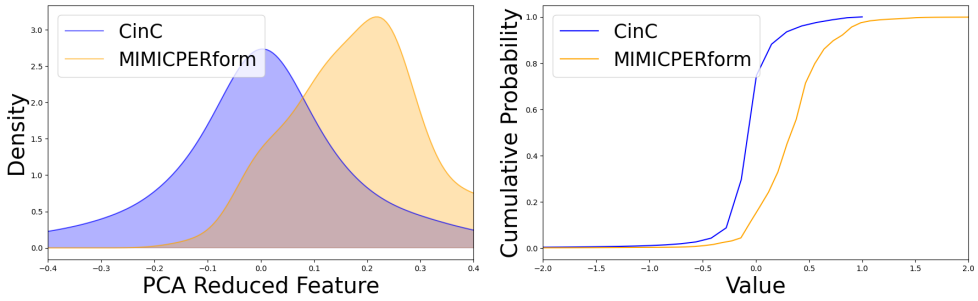
MIMICPERform PPG Dataset. This dataset contains PPG recordings from the same patients and timeframes as the MIMICPERform ECG set described above. While it does not include ECG signals, which are the focus of this paper, we use it to test ADAPTOOD's ability to adapt across biosignal modalities.

CODEtest Dataset. This dataset is the test version of the Clinical Outcomes in Digital Electrocardiology (CODE) study [220, 221] and includes 827 ECG recordings, each consisting of 12 leads sampled at 400 Hz, with 4 096 data points per lead. We classify the samples into healthy and non-healthy groups, and the recordings in this dataset have been annotated by cardiologists, medical students and others. We use the gold standard annotation provided in the original source [221], obtained through a structured process involving two cardiologists with disagreements resolved by a senior specialist.

5.2.2 OOD Scenarios

We chose all datasets so that they differ in varying aspects compared to the pre-training data. The shift details are listed in Table 5.1.

MIT-BIH. Compared to the pre-training data, this introduces sensor shifts due to different acquisition devices, population shifts because of different patient cohorts, and temporal shifts from changes in sampling contexts. To visually illustrate these differences, all features are scaled to a common range, and PCA reduces dimensionality to a single



(a) KDE Plot (b) CDF Plot

Figure 5.4: Distribution shift visualisations between the pre-training data and the ECG data of MIMICPERform.

component (1D) for direct comparison. Figure 5.2a presents a Kernel Density Estimate (KDE) plot, where filled areas indicate the highest concentration of values, while Figure 5.2b shows a Cumulative Distribution Function (CDF) plot, illustrating cumulative probabilities across values. The separation between the curves in both plots visually confirms the presence of distribution shifts.

PTB-DB. This dataset introduces population shifts from differing clinical sources, sensor shifts due to changes in measurement equipment, and temporal shifts in signal dynamics. For its OOD severity, we use the same visualisation strategy as before. As per the KDE plot of Figure 5.3a, both datasets have similar unimodal distributions, but the PTB-DB (orange) is slightly shifted and narrower compared to the pre-training data (blue), indicating changes in feature spread and central tendency. The CDF plot in Figure 5.3b further confirms this through differences in probability mass accumulation.

MIMICPERform ECG. This dataset presents population (age and condition differences), label (annotations), sensor (monitoring setups), and temporal (clinical context) shifts. To illustrate these, Figure 5.4a shows the KDE plot: MIMICPERform ECG data (orange) have a noticeable rightward shift and broader distribution than PhysioNet CinC (blue), reflecting changes in central tendency and feature variability. Unlike earlier datasets, MIMICPERform's density peak is lower and wider, indicating a more heterogeneous feature distribution. In the CDF plot of Figure 5.4b, the shift is further emphasised: the curve for MIMICPERform rises more gradually and begins accumulating mass at higher values compared to the pre-training dataset. Therefore, the MIMICPERform ECG signals occupy a different feature space, challenging model generalisation.

MIMICPERform PPG. This introduces a modality shift, since the ECG pre-trained model is now applied to PPG data, which have different physiological properties as they measure blood volume changes at the skin surface using light rather than electrical cardiac activity. This is the most challenging shift, and although ECG models are not expected to perform well on such fundamentally different signals, this setup provides insights into the limits of model generalisation across biosignal types. For this reason, we do not include

plots comparing ECG and PPG, as their underlying signal properties differ fundamentally, and comparisons would be misleading without a shared feature representation.

CODEtest Dataset. Further to population, sensor, and temporal shifts, this dataset also introduces a dimensionality shift by increasing the number of input channels. This is in contrast to previous cases, making it valuable for assessing ADAPTOOD's robustness.

5.3 Experimental Setup

In this section, we describe the model architecture, the baselines, and the ablation studies selected for our experiments.

5.3.1 Model Architecture

We adopt a 1D convolutional neural network (CNN) as our pre-trained model architecture due to its strong inductive bias for temporal signal processing, computational efficiency, and widespread use in ECG analysis, where it can capture local patterns while remaining scalable to long time series.

The architecture starts with an input layer accepting vectors in the shape of the data, followed by several stacked Conv1D layers with increasing filters (32, 64, 128, 256, 512) and a kernel size of 5, each using ReLU activation and the same padding. Subsequently, MaxPooling1D layers reduce the spatial dimension after each convolution, and Dropout layers are inserted after higher-capacity layers to prevent overfitting. After the final convolution, the output is flattened and passed through two fully connected dense layers with 64 and 32 neurons respectively, both using the ReLU activation function, followed by a final dense output layer with a single neuron and a sigmoid activation function for binary classification. The model is trained in a supervised manner for 20 epochs to ensure convergence, and compiled using the settings chosen dynamically by the hyperaparameter tuner of Section 5.1.6.

5.3.2 Baselines & Ablations

To evaluate our approach, we establish some key alternatives.

Transfer Learning. This reflects a conventional transfer learning setup and operates by fine-tuning a pre-trained model to extract features from the input data. A subset of the model's earlier layers serve as a feature extractor, with their weights frozen to preserve learned representations. The remaining layers are trainable, and the model is compiled using the original optimiser and loss function.

Supervised Learning. This works in a supervised fashion using a deep convolutional neural network with the same model layers as ADAPTOOD, described in Section 5.3.1.

It is compiled using a cross-entropy loss and the Adam optimiser [179]. Across all experiments, this baseline has access to the same total amount of samples as the alternatives, representing a fair comparison of how the model would perform when there is a only a niche dataset annotated.

Feature-Based Domain Adaptation. This illustrates a feature-based domain adaptation approach, which builds a shared feature representation to correct the difference between the source and target distributions. The task is then learned in this encoded space. For this baseline we use the ADAPT toolbox [222], and specifically its PRED strategy [223] that is widely used and frequently cited. This first trains a model on the source domain and uses its predictions on the target data as additional input features. A second model is then trained on the labelled target data, augmented with these features. For a fair comparison, the source model uses the same architecture as ADAPTOOD.

Instance-Based Domain Adaptation. Further to the feature-based domain adaptation baseline seen above, we also use an instance-based domain adaptation alternative. Instead of researching common features, this approach focuses on reweighting training data in order to correct the difference between the source and target distributions. To implement this, we rely on the widely used nearest neighbours weighting strategy, which reweights the source instances according to their number of neighbours in the target dataset [224]. During training, this reweighting involves multiplying the loss of each training instance by a positive weight. For a fair comparison, the estimator used to learn the task uses the same architecture as ADAPTOOD's model.

Ablation using only the Hellinger Distance. To enhance our comprehension of which metric is more effective for estimating uncertainty through the calculation of distribution-level divergence, we initially conduct an ablation study employing solely the Hellinger distance. All other elements of ADAPTOOD are retained, thus ensuring that any observed performance improvements can be attributed exclusively to the Hellinger distance.

Ablation using only the Mahalanobis Distance. Likewise, a separate ablation study is conducted using exclusively the Mahalanobis distance. Although our selected metrics are complementary, the sole use of the Mahalanobis distance in this ablation highlights its distinct advantages over using the Hellinger distance or both.

Ablation without Hyperparameter Tuning. To better understand the impact of ADAPTOOD's additional components, this ablation removes its hyperparameter tuning functionalities. All other aspects remain unchanged, including the OOD adaptation mechanism and the LoRA module. This setup isolates the performance gains due to improved hyperparameter optimisation. To this end, the relevant hyperparameters are fixed and set to use the Adam optimiser [179] and its default learning rate of 1e−3.

Ablation without LoRA. To assess the effect of low-rank adaptation on ADAPTOOD's performance, we run a set of experiments by disabling it. These examine how

it affects the adapted model in terms of size, memory footprint, and parameter count, potentially introducing overfitting relative to our full proposal.

5.4 Evaluation Results

In Tables 5.2, 5.4, and 5.5 we present the results of our evaluation. For greater consistency, we conduct three repeated trials per experiment and report the average values obtained. ADAPTOOD is found to achieve the best performance across scenarios.

5.4.1 Key Findings

Under population, sensor, and temporal shifts, which occur for both MIT-BIH and PTB-DB as per Section 5.2.2, ADAPTOOD achieves strong improvements by guiding adaptation based on OOD severity.

In the MIT-BIH data, these shifts cause transfer learning to perform very closely to supervised learning in accuracy (0.910 vs 0.906) and precision (0.891 vs 0.914). Yet, our OOD severity-based mechanism improves performance across metrics, including in accuracy (0.953) and precision (0.930). While the sensitivity of these baselines remains near chance level (0.595 and 0.499), ADAPTOOD achieves 0.820 by tailoring its adaptation to OOD severity, leading to more generalisable classification. The feature-based domain adaptation baseline shows improvements, but ADAPTOOD still outperforms it in all cases, including in accuracy (0.953 vs 0.940), precision (0.930 vs 0.921), and recall (0.900 vs 0.861). Specificity is inflated for supervised learning compared to ADAPTOOD (0.991 vs 0.980), but this results from the baseline's tendency to underpredict the positive class, leading to unbalanced performance that is not desirable in healthcare.

Similar results are observed in the PTB-DB data, where transfer learning and supervised learning perform nearly identically, with no effective gains over supervised learning. The domain adaptation cases, which are more advanced by design, performs better, but ADAPTOOD still outperforms all. For instance, in accuracy (0.922) it achieves improvements of 11.9% over the transfer learning (0.803), 10.4% over the supervised learning (0.818), 4.7% over the feature-based domain adaptation (0.875), and 4.9% over the instance-based domain adaptation baselines. A similar trend is recorded in the recall (0.899 vs 0.797 vs 0.728 vs 0.836 vs 0.835), while in precision it achieves a 12.9% increase compared to transfer learning (0.901 vs 0.772), showing the benefits of its OOD-based mechanism.

When, in addition to the above shifts, there is a task and label shift also introduced, the OOD scenario becomes even more severe. This is the case for the ECG MIMICER-formTT dataset, as per Section 5.2.2, but ADAPTOOD continues to demonstrate superior generalisation. It achieves an accuracy of 0.942, significantly higher than both the transfer

Table 5.2: Evaluation results from testing our approach on OOD ECG classification tasks on baselines and our ADAPTOOD method.

Method	Accuracy	Precision	Recall	Sensitivity	Specificity	F1 Score
	ECG MIT-BIH					
Transfer Learning	0.910 ± 0.007	0.891 ± 0.045	0.786 ± 0.055	0.595 ± 0.132	0.976 ± 0.024	0.818 ± 0.030
Supervised Learning	0.906 ± 0.019	0.914 ± 0.009	0.745 ± 0.056	0.499 ± 0.114	**0.991 ± 0.004**	0.793 ± 0.056
Feature-Based DA	0.940 ± 0.002	0.921 ± 0.003	0.861 ± 0.011	0.740 ± 0.023	0.982 ± 0.003	0.887 ± 0.006
Instance-Based DA	0.936 ± 0.004	0.921 ± 0.003	0.847 ± 0.019	0.710 ± 0.039	0.983 ± 0.003	0.878 ± 0.012
ADAPTOOD	**0.953 ± 0.002**	**0.930 ± 0.008**	**0.900 ± 0.005**	**0.820 ± 0.013**	0.980 ± 0.005	**0.914 ± 0.003**
	ECG PTB-DB					
Transfer Learning	0.803 ± 0.033	0.772 ± 0.030	0.797 ± 0.023	0.811 ± 0.096	0.784 ± 0.139	0.769 ± 0.015
Supervised Learning	0.818 ± 0.007	0.786 ± 0.007	0.728 ± 0.037	0.924 ± 0.028	0.531 ± 0.102	0.745 ± 0.028
Feature-Based DA	0.875 ± 0.000	0.851 ± 0.013	0.836 ± 0.038	0.920 ± 0.045	0.753 ± 0.120	0.838 ± 0.012
Instance-Based DA	0.873 ± 0.003	0.843 ± 0.007	0.835 ± 0.025	0.918 ± 0.024	0.753 ± 0.074	0.837 ± 0.011
ADAPTOOD	**0.922 ± 0.013**	**0.901 ± 0.019**	**0.899 ± 0.015**	**0.948 ± 0.014**	**0.852 ± 0.019**	**0.900 ± 0.016**
	ECG MIMICPERformTT					
Transfer Learning	0.882 ± 0.025	0.895 ± 0.015	0.882 ± 0.025	0.950 ± 0.045	0.813 ± 0.095	0.881 ± 0.026
Supervised Learning	0.862 ± 0.033	0.874 ± 0.035	0.862 ± 0.033	0.920 ± 0.090	0.803 ± 0.075	0.861 ± 0.033
Feature-Based DA	0.867 ± 0.003	0.869 ± 0.003	0.867 ± 0.003	0.827 ± 0.020	**0.907 ± 0.020**	0.866 ± 0.003
Instance-Based DA	0.842 ± 0.008	0.851 ± 0.008	0.842 ± 0.008	0.923 ± 0.010	0.760 ± 0.010	0.841 ± 0.008
ADAPTOOD	**0.942 ± 0.003**	**0.944 ± 0.002**	**0.942 ± 0.003**	**0.980 ± 0.010**	0.903 ± 0.015	**0.942 ± 0.003**
	PPG MIMICPERformTT					
Transfer Learning	0.917 ± 0.008	0.929 ± 0.006	0.917 ± 0.008	**1.000 ± 0.000**	0.833 ± 0.015	0.916 ± 0.008
Supervised Learning	0.905 ± 0.015	0.920 ± 0.011	0.905 ± 0.015	**1.000 ± 0.000**	0.810 ± 0.030	0.904 ± 0.016
Feature-Based DA	0.942 ± 0.003	0.946 ± 0.004	0.942 ± 0.003	0.990 ± 0.010	0.893 ± 0.005	0.942 ± 0.003
Instance-Based DA	0.945 ± 0.005	0.948 ± 0.007	0.945 ± 0.005	0.983 ± 0.015	0.907 ± 0.005	0.945 ± 0.005
ADAPTOOD	**0.975 ± 0.010**	**0.976 ± 0.009**	**0.975 ± 0.010**	0.995 ± 0.005	**0.953 ± 0.025**	**0.975 ± 0.010**
	ECG CODEtest					
Transfer Learning	0.873 ± 0.007	0.825 ± 0.057	0.757 ± 0.050	0.565 ± 0.143	0.949 ± 0.043	0.773 ± 0.013
Supervised Learning	0.861 ± 0.019	0.825 ± 0.027	0.669 ± 0.034	0.361 ± 0.069	**0.978 ± 0.001**	0.706 ± 0.035
Feature-Based DA	0.894 ± 0.006	0.803 ± 0.007	0.778 ± 0.025	0.611 ± 0.063	0.945 ± 0.012	0.789 ± 0.016
Instance-Based DA	0.882 ± 0.046	0.765 ± 0.082	0.734 ± 0.043	0.528 ± 0.054	0.939 ± 0.033	0.746 ± 0.058
ADAPTOOD	**0.922 ± 0.018**	**0.875 ± 0.053**	**0.807 ± 0.052**	**0.641 ± 0.093**	0.974 ± 0.011	**0.836 ± 0.053**

learning (0.882) and domain adaptation (0.867 and 0.842) baselines. This pattern extends to the F1 score, where ADAPTOOD reaches 0.942 compared to 0.881, 0.866, and 0.841 for the respective baselines. These results further demonstrate the robustness of the OOD severity-based mechanism in the presence of compound distributional shifts, maintaining balanced fine-tuning.

To further test ADAPTOOD, we examine a different modality, specifically the photoplethysmography (PPG) recordings of the MIMICPERform PPG dataset (see Section 5.2.1). This introduces a modality shift in addition to the shifts considered in previous scenarios, representing an edge case, as PPG signals exhibit different physiological properties compared to the ECG data used for pre-training. In this case, transfer learning brings

a marginal improvement of only 1% in accuracy over the supervised baseline (0.917 vs 0.905), while the instance-based domain adaptation offers a better solution that results in 4% higher accuracy (0.945 vs 0.905). Yet, ADAPTOOD outperforms all by reaching 0.975 in accuracy, while similar results are recorded in the other examined metrics too, like the precision and the recall. Although fine-tuning is rarely the go-to solution when adapting models across modalities and ADAPTOOD is not designed for such cases, this experiment demonstrates its ability to also handle more severe OOD scenarios.

In the case of the CODEtest dataset, which introduces a dimensionality shift by increasing the number of input channels, the results are also interesting. That's because it was originally recorded using 12-lead configurations of ECG biosignals, rather than single-lead like the previous datasets. The transfer learning, supervised learning, feature-based domain adaptation and instance-based domain adaptation baselines reach values of 0.873, 0.861, 0.894, and 0.882 in accuracy, respectively, but ADAPTOOD outperforms all with a value of 0.922. ADAPTOOD processes this dataset to make it compatible with the pre-trained model by repurposing it to use a single channel, but this does not affect the fact that it represents an even more challenging OOD case than previous ones. Our solution's improvements also extend to other metrics, including in precision (0.875 vs less than 0.825 for the baselines) and F1 score (0.836 vs less than 0.789 for the baselines). OOD generalisation proves challenging for alternatives that do not account for its severity, but ADAPTOOD is consistent in its improvements.

5.4.2 Robustness Across Levels of Distribution Shift Severity

Further to ADAPTOOD's evaluation results in absolute terms, it is also useful to observe how the numerical magnitude of the distribution shift affects the performance of the proposed approach and the baselines, and how this evolves as severity increases.

Table 5.1, seen previously, characterises the types of distribution shifts observed between the pre-training data and each downstream dataset, and Figures 5.2, 5.3, and 5.4 show the distribution shift severity visually. To complement these, we now categorise the datasets of our case studies into five severity levels:

- Low Severity: ECG CODEdata with a Hellinger distance of 2.24
- Medium Severity: ECG MIT-BIH with a Hellinger distance of 4.62
- High Severity: ECG PTB-DB with a Hellinger distance of 5.47
- Very High Severity: ECG MIMICPERformTT with a Hellinger distance of 6.32
- Extreme Severity: PPG MIMICPERformTT with a Hellinger distance of 8.30

Table 5.3: Robustness analysis across varying levels of distribution shift severity.

Method	Low Severity	Medium Severity	High Severity	Very High Severity	Extreme Severity
	Accuracy				
Transfer Learning	0.873 ± 0.007	0.910 ± 0.007	0.803 ± 0.033	0.882 ± 0.025	0.917 ± 0.008
Supervised Learning	0.861 ± 0.019	0.906 ± 0.019	0.818 ± 0.008	0.862 ± 0.033	0.905 ± 0.015
Feature-Based DA	0.894 ± 0.006	0.940 ± 0.003	0.875 ± 0.000	0.867 ± 0.003	0.942 ± 0.003
Instance-Based DA	0.882 ± 0.046	0.936 ± 0.004	0.873 ± 0.003	0.842 ± 0.008	0.945 ± 0.005
ADAPTOOD	**0.922 ± 0.018**	**0.953 ± 0.002**	**0.922 ± 0.013**	**0.942 ± 0.003**	**0.975 ± 0.010**
% Improvement over Best Baseline	3.13%	1.35%	5.33%	6.80%	3.18%
	F1 Score				
Transfer Learning	0.773 ± 0.013	0.818 ± 0.030	0.769 ± 0.015	0.881 ± 0.026	0.916 ± 0.008
Supervised Learning	0.706 ± 0.035	0.793 ± 0.056	0.745 ± 0.028	0.861 ± 0.033	0.904 ± 0.016
Feature-Based DA	0.789 ± 0.016	0.887 ± 0.006	0.838 ± 0.012	0.866 ± 0.003	0.942 ± 0.003
Instance-Based DA	0.746 ± 0.058	0.878 ± 0.012	0.837 ± 0.011	0.841 ± 0.008	0.945 ± 0.005
ADAPTOOD	**0.836 ± 0.053**	**0.914 ± 0.003**	**0.900 ± 0.016**	**0.942 ± 0.003**	**0.975 ± 0.010**
% Improvement over Best Baseline	5.96%	3.12%	7.40%	6.93%	3.18%

This categorisation allows us to examine the robustness of ADAPTOOD under progressively more challenging OOD conditions, complementing the qualitative shift taxonomy presented earlier.

Table 5.3 reports the accuracy and F1 score achieved by all baselines and ADAPTOOD across these severity levels, together with the percentage improvement of ADAPTOOD over the best-performing alternative in each case. Under low severity, where the target data are relatively similar to the pre-training distribution, all methods perform strongly. Nevertheless, ADAPTOOD achieves an accuracy of 0.922 compared to the strongest baseline at 0.894, corresponding to a 3.13% improvement. In terms of F1 score, the gain is more pronounced, improving from 0.789 to 0.836 (5.96% improvement). This indicates that even under mild shifts, uncertainty-guided adaptation provides better calibration between precision and recall. At medium severity levels, the accuracy improves by a lower amount, specifically from 0.940 to 0.953 (1.35%), while the F1 score increases from 0.887 to 0.914 (3.12%).

As severity increases to high levels, the relative advantages of ADAPTOOD become more substantial. Under high severity, the accuracy rises from 0.875 to 0.922 (5.33%), and the F1 score from 0.838 to 0.900 (7.40%). These gains suggest that as the divergence between source and target distributions grows, conventional transfer learning and domain adaptation approaches struggle to maintain balanced performance, whereas the severity-aware mechanism in ADAPTOOD enables more effective calibration. The benefits are further amplified at very high severity. In this case, ADAPTOOD achieves 0.942 accuracy compared to the strongest baseline at 0.882 (6.80% improvement), and 0.942 in the F1

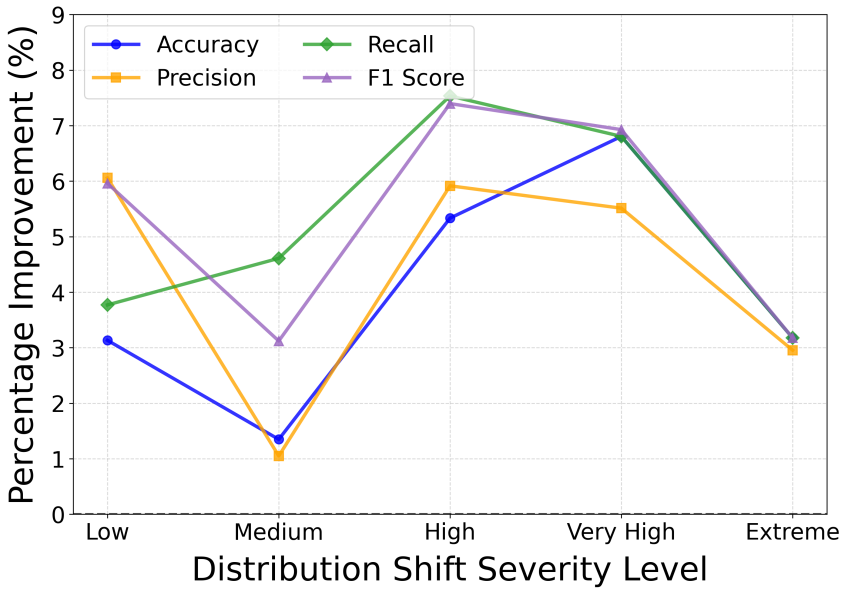


Figure 5.5: Percentage improvement of ADAPTOOD across varying levels of distribution shift severity, compared to the best-performing alternative in each case.

score compared to 0.881 in the strongest baseline (6.93% improvement). This demonstrates that when the shift is substantial but still within the same physiological modality (ECG-to-ECG), uncertainty-guided adaptation meaningfully stabilises performance.

At the extreme severity level, corresponding to cross-modality adaptation from an ECG pre-training signal to a PPG signal, ADAPTOOD maintains the highest absolute performance (0.975 accuracy and 0.975 F1). However, the relative improvement over the best baseline reduces to 3.18% for both metrics.

Figure 5.5 provides further insight into the observed model behaviour. It shows that percentage improvements generally increase from low to high and very high severities, particularly for the recall and the F1 score, before dropping at the extreme level. Notably, the recall and the F1 score exhibit the largest relative gains at high severity, exceeding 7%, indicating that ADAPTOOD is especially effective under pronounced distribution shifts. Precision improvements follow a similar but slightly smoother trend, while accuracy improvements peak at the very high level. The drop in relative improvement at the extreme severity level is consistent with the nature of the shift. Unlike the other cases, this scenario involves a modality change (ECG to PPG), representing a fundamentally different signal space rather than a variation within the same modality. Under such extreme divergence, the adaptation challenge becomes qualitatively different. Consequently, while ADAPTOOD still achieves the best absolute performance, the relative margin narrows. This behaviour further supports our central hypothesis: severity-aware adaptation yields the greatest relative benefit when shifts are substantial but structurally related to the source domain, whereas cross-modality shifts introduce additional complexities that limit the proportional gains.

Table 5.4: Ablation study comparing variants of our approach without HP tuning and without LoRA with the full method.

Method	Accuracy	Precision	Recall	Sensitivity	Specificity	F1 Score
	ECG MIT-BIH					
Hellinger Dst	0.930 ± 0.023	0.912 ± 0.011	0.832 ± 0.076	0.682 ± 0.157	0.981 ± 0.006	0.861 ± 0.058
Mahalanobis Dst	0.951 ± 0.007	0.929 ± 0.012	0.894 ± 0.016	0.807 ± 0.030	**0.981 ± 0.004**	0.910 ± 0.014
ADAPTOOD	**0.953 ± 0.002**	**0.930 ± 0.008**	**0.900 ± 0.005**	**0.820 ± 0.013**	0.980 ± 0.005	**0.914 ± 0.003**
	ECG PTB-DB					
Hellinger Dst	0.825 ± 0.045	0.839 ± 0.024	0.719 ± 0.133	0.950 ± 0.058	0.488 ± 0.324	0.722 ± 0.129
Mahalanobis Dst	0.875 ± 0.045	0.871 ± 0.034	0.800 ± 0.089	**0.964 ± 0.007**	0.636 ± 0.185	0.819 ± 0.083
ADAPTOOD	**0.922 ± 0.013**	**0.901 ± 0.019**	**0.899 ± 0.015**	0.948 ± 0.014	**0.852 ± 0.019**	**0.900 ± 0.016**
	ECG MIMICPERformTT					
Hellinger Dst	0.933 ± 0.028	0.938 ± 0.022	0.933 ± 0.028	0.980 ± 0.010	0.887 ± 0.060	0.933 ± 0.028
Mahalanobis Dst	0.937 ± 0.005	0.942 ± 0.003	0.937 ± 0.005	**0.990 ± 0.010**	0.883 ± 0.020	0.937 ± 0.005
ADAPTOOD	**0.942 ± 0.003**	**0.944 ± 0.002**	**0.942 ± 0.003**	0.980 ± 0.010	**0.903 ± 0.015**	**0.942 ± 0.003**
	PPG MIMICPERformTT					
Hellinger Dst	0.958 ± 0.023	0.962 ± 0.020	0.958 ± 0.023	0.995 ± 0.005	0.920 ± 0.050	0.958 ± 0.023
Mahalanobis Dst	0.965 ± 0.023	0.968 ± 0.020	0.965 ± 0.023	1.000 ± 0.000	0.930 ± 0.045	0.965 ± 0.023
ADAPTOOD	**0.975 ± 0.010**	**0.976 ± 0.009**	**0.975 ± 0.010**	0.995 ± 0.005	**0.953 ± 0.025**	**0.975 ± 0.010**
	ECG CODEtest					
Hellinger Dst	0.914 ± 0.027	0.914 ± 0.023	0.799 ± 0.119	0.622 ± 0.252	**0.977 ± 0.020**	0.824 ± 0.100
Mahalanobis Dst	0.912 ± 0.021	0.873 ± 0.035	**0.824 ± 0.072**	**0.683 ± 0.144**	0.965 ± 0.003	**0.843 ± 0.057**
ADAPTOOD	**0.922 ± 0.018**	**0.875 ± 0.053**	0.807 ± 0.052	0.641 ± 0.093	0.974 ± 0.011	0.836 ± 0.053

Overall, the findings from Table 5.3 and Figure 5.5 confirm that ADAPTOOD delivers consistent and scalable improvements across a spectrum of distribution shift severities. By explicitly quantifying OOD severity and adjusting the adaptation strategy accordingly, the framework maintains robust performance from mild to extreme distribution shifts, demonstrating good generalisability across settings.

5.4.3 Ablation Studies

While ADAPTOOD's uncertainty-guided transfer learning system serves as its core engine, both the choice of uncertainty metrics and the choice of its supporting components is imperative towards its success in adapting under varying OOD severities.

Uncertainty Metric Comparisons

To better understand which of the metrics used in our proposal, seen in Section 5.1.3, is the most effective in capturing uncertainty related to OOD severity, we conduct an ablation study with three distinct sets of experiments: one utilising only the Hellinger distance metric, another using only the Mahalanobis distance metric, and a third applying

Table 5.5: Ablation study results, comparing uncertainty metric variants of our approach with the full method.

Method	Accuracy	Precision	Recall	Sensitivity	Specificity	F1 Score
	ECG MIT-BIH					
w/o HP Tuning	0.945 ± 0.004	0.919 ± 0.026	0.887 ± 0.025	0.798 ± 0.061	0.976 ± 0.015	0.900 ± 0.009
w/o LoRA	0.949 ± 0.004	0.915 ± 0.011	**0.904 ± 0.009**	**0.835 ± 0.018**	0.973 ± 0.006	0.910 ± 0.008
ADAPTOOD	**0.953 ± 0.002**	**0.930 ± 0.008**	0.900 ± 0.005	0.820 ± 0.013	**0.980 ± 0.005**	**0.914 ± 0.003**
	ECG PTB-DB					
w/o HP Tuning	0.882 ± 0.008	0.869 ± 0.009	0.822 ± 0.020	0.952 ± 0.014	0.691 ± 0.046	0.840 ± 0.014
w/o LoRA	0.898 ± 0.008	0.885 ± 0.010	0.851 ± 0.017	**0.954 ± 0.011**	0.747 ± 0.037	0.865 ± 0.012
ADAPTOOD	**0.922 ± 0.013**	**0.901 ± 0.019**	**0.899 ± 0.015**	0.948 ± 0.014	**0.852 ± 0.019**	**0.900 ± 0.016**
	ECG MIMICPERformTT					
w/o HP Tuning	0.913 ± 0.020	0.923 ± 0.012	0.913 ± 0.020	**0.983 ± 0.015**	0.843 ± 0.050	0.913 ± 0.021
w/o LoRA	0.918 ± 0.020	0.925 ± 0.015	0.918 ± 0.020	0.977 ± 0.005	0.860 ± 0.045	0.918 ± 0.021
ADAPTOOD	**0.942 ± 0.003**	**0.944 ± 0.002**	**0.942 ± 0.003**	0.980 ± 0.010	**0.903 ± 0.015**	**0.942 ± 0.003**
	PPG MIMICPERformTT					
w/o HP Tuning	0.953 ± 0.025	0.958 ± 0.022	0.953 ± 0.025	1.000 ± 0.000	0.907 ± 0.050	0.953 ± 0.025
w/o LoRA	0.952 ± 0.005	0.956 ± 0.005	0.952 ± 0.005	1.000 ± 0.000	0.903 ± 0.010	0.952 ± 0.005
ADAPTOOD	**0.975 ± 0.010**	**0.976 ± 0.009**	**0.975 ± 0.010**	0.995 ± 0.005	**0.953 ± 0.025**	**0.975 ± 0.010**
	ECG CODEtest					
w/o HP Tuning	0.906 ± 0.003	0.869 ± 0.024	0.753 ± 0.019	0.527 ± 0.039	0.978 ± 0.007	0.794 ± 0.020
w/o LoRA	0.916 ± 0.024	**0.927 ± 0.017**	0.792 ± 0.081	0.593 ± 0.173	**0.990 ± 0.011**	0.831 ± 0.063
ADAPTOOD	**0.922 ± 0.018**	0.875 ± 0.053	**0.807 ± 0.052**	**0.641 ± 0.093**	0.974 ± 0.011	**0.836 ± 0.053**

our full approach. The results of these experiments can be found in Table 5.4 and allow us to isolate the performance contributions of each metric. While the Mahalanobis distance demonstrates superior performance compared to Hellinger, our full approach that integrates both metrics consistently outperforms the others across evaluated criteria. For instance, in the PTB-DB data it reaches an accuracy of 0.922, compared to just 0.825 when using the Hellinger distance alone or 0.875 when using the Mahalanobis distance alone. Similarly, in the ECG MIMICPERformTT data, ADAPTOOD reaches a recall value of 0.942, while the Hellinger distance ablation has a lower value of 0.933. While the differences between the ablations and the full proposal are not significant in the case of switching the metrics used, it is clear that combining these metrics allows for more consistent and robust uncertainty estimation across varying tasks and datasets.

Supporting Mechanism Comparisons

Further to examining ADAPTOOD's uncertainty-based ablations, we also conduct a comprehensive study to understand the importance of its supporting components for effective adaptation. We observe that both the hyperparameter tuning and the LoRA integration

contribute to its performance gains on several datasets.

The hyperparameter (HP) tuning proves particularly useful in most cases. In ECG PTB-DB, removing it lowers ADAPTOOD's accuracy by 4% (0.922 vs 0.882), while this is also confirmed through other metrics like the precision (0.901 vs 0.869) and the recall (0.899 vs 0.822) that also get reduced. A similar drop is noted in ECG MIMICPERformTT, where the ablation without HP tuning shows a 3% drop in accuracy (0.942 vs 0.913). However, the HP tuning module is not as significant in some other cases. For instance, its removal in the MIT-BIH data leads to less than 1% drop in accuracy (0.953 vs 0.945), while having no impact in metrics like the specificity (0.980 vs 0.976). This shows that while selective layer unfreezing primarily drives adaptation, HP tuning plays a key supporting role when the pre-trained model is less suited to the OOD levels.

The significance of the LoRA module shows resembling behaviour: some of the ablations that do not use it suffer from negligible performance degradation, while for others it proves crucial. In the ECG MIMICPERformTT data, its removal decreases the accuracy and F1 score by 2.9% (both 0.942 vs 0.913) and the specificity by as much as 6% (0.903 vs 0.843). In contrast, in the ECG CODEtest data it results in just a 0.6% decrease in accuracy (0.922 vs 0.916), while surprisingly improving precision (0.875 vs 0.927) and specificity (0.990 vs 0.974) compared to the full ADAPTOOD model. Yet, these improvements are likely due to noise or dataset-specific variance and, according to the other results of Table 5.4, they do not generalise. Therefore, using the full ADAPTOOD model, including the LoRA module, is recommended for consistent and robust adaptation under varying OOD severity scenarios.

5.4.4 Model Efficiency

ADAPTOOD is computationally efficient, and this is achieved through its uncertainty-based adaptation approach and the use of low-rank model updates that enable parameter-efficient adaptation. As per Table 5.6, across all evaluated datasets, the full ADAPTOOD system leads to a substantial reduction in both total parameter count and memory footprint compared to the ablation without LoRA. For instance, in the MIT-BIH dataset, ADAPTOOD's final model requires only 2 256 225 total parameters (8.61 MB) versus 7 647 781 parameters (29.17 MB) for the version without LoRA. Similar trends are observed for the PTB-DB (2 256 225 vs. 7 647 781 parameters; 8.61 MB vs. 29.17 MB), ECG MIMICPERformTT (21 392 737 vs. 26 784 293 parameters; 81.61 MB vs. 102.17 MB), PPG MIMICPERformTT (21 392 737 vs. 26 784 293 parameters; 81.61 MB vs. 102.17 MB), and CODEtest (14 806 369 vs. 20 197 925 parameters; 56.48 MB vs. 77.05 MB) data. Moreover, ADAPTOOD updates only the minimal set of parameters necessary, rather than fine-tuning the full model. This leads to computational cost and storage savings. Its general design also emphasises adaptive updates: uncertainty-aware

Table 5.6: Model size in variants of our approach without HP tuning, without LoRA, and with the full ADAPTOOD method.

Method	Parameter Count	Parameter Size
	ECG MIT-BIH	
w/o HP Tuning	2,256,225	8.61 MB
w/o LoRA	7,647,781	29.17 MB
ADAPTOOD	**2,256,225**	**8.61 MB**
	ECG PTB-DB	
w/o HP Tuning	2,256,225	8.61 MB
w/o LoRA	7,647,781	29.17 MB
ADAPTOOD	**2,256,225**	**8.61 MB**
	ECG MIMICPERformTT	
w/o HP Tuning	21,392,737	81.61 MB
w/o LoRA	26,784,293	102.17 MB
ADAPTOOD	**21,392,737**	**81.61 MB**
	PPG MIMICPERformTT	
w/o HP Tuning	21,392,737	81.61 MB
w/o LoRA	26,784,293	102.17 MB
ADAPTOOD	**21,392,737**	**81.61 MB**
	ECG CODEtest	
w/o HP Tuning	14,806,369	56.48 MB
w/o LoRA	20,197,925	77.05 MB
ADAPTOOD	**14,806,369**	**56.48 MB**

adjustments mean that minimal updates are made for high-confidence cases, while more extensive fine-tuning is performed for low-confidence cases. Thus, it is a lightweight yet powerful framework for real-world cases.

5.5 Conclusions

This chapter presented ADAPTOOD, a framework introducing dynamic and uncertainty-aware fine-tuning of pre-trained models for ECG biosignals under OOD conditions. It directly addresses the complex and varied distribution shifts seen in real-world ECGs, arising from differences in populations, sensors, domains, labels, and temporal contexts. By fine-tuning only when necessary and adjusting critical model layers, it avoids overfitting and reduces unnecessary computation. Its adaptive strategy leverages uncertainty to guide selective layer training, applies LoRA to update relevant components, and uses automated hyperparameter optimisation to adjust model settings effectively.

The ADAPTOOD method outperforms alternatives across a broad range of OOD scenarios, showing consistent gains in robustness, calibration, and generalisation. Its ablations perform strongly as well, which is important as it shows that it can also be a reliable option in settings with varying scalability, complexity, or real-time deployment requirements. A key contributor to this performance is ADAPTOOD's ability to adapt the degree of fine-tuning based on the severity of distribution shifts, indicating the benefit of uncertainty-guided adaptation over static protocols. These results position ADAPTOOD as a promising and reliable approach for out-of-distribution model fine-tuning cases.

Chapter 6

Discussion & Conclusions

This thesis presented original research on how machine learning models can become more data-efficient, adaptive, and generalisable across diverse healthcare use cases through uncertainty-aware learning. As discussed earlier, this is crucial since sensor-based datasets are frequently sparse (irregularly-sampled) or unlabelled when collected in real-world environments. Additionally, ML models often exhibit limited generalisability under out-of-distribution conditions, which limits the use of real-world time series data in deep learning tasks. To address these challenges, this thesis presented three primary contributions that are summarised in this chapter, which also reflects on their implications for ML applications in real-world data and suggests future research directions.

6.1 Summary of Contributions

This section revisits the research questions introduced in Chapter 1, reflecting on their significance and how they have shaped the study, and also summarises the key contributions of the thesis, highlighting the most important findings.

6.1.1 Modelling Sequence-to-Sequence Solutions for Sparse Time Series

Research Question 1. *How can we design sequence-to-sequence models that effectively capture temporal dependencies in sparse healthcare time series while providing reliable and informative uncertainty estimates?*

Contribution 1. In Chapter 3 we introduced SQUIREDL, a novel uncertainty-aware sequence-to-sequence prediction method for sparse healthcare time series. The motivation behind the development of this approach was that most machine learning models assume regular spacing in time series, but this is often unrealistic in healthcare, where missing data is common, limiting model performance. Healthcare time series are rarely complete, typically due to missed sensor recordings or wrong sensing. The integration of uncertainty

into sequence-to-sequence models for irregular time series remains underdeveloped, and this negatively impacts model performance.

To address these challenges, SQUIREDL enhances the state-of-the-art evidential regression framework, widely used for uncertainty estimation, to handle missing data. After imputing data with an Akima spline-based method, the loss function of evidential regression was modified by assigning different weights to imputed and observed data points, providing more reliable uncertainty estimates. Additionally, SQUIREDL examines various metrics to assess the success of uncertainty estimations in sequence-to-sequence predictions, offering a robust and reliable way to evaluate models in a medical setting.

We demonstrated SQUIREDL in clinical applications, starting with continuous glucose monitoring, where sequence-to-sequence prediction was used to estimate hypoglycemia risk from glucose sensor readings. In this case, SQUIREDL achieved 30% higher accuracy in capturing the ground truth risk values, showing consistent improvements in both uncertainty-aware and accuracy-based metrics. Similarly, in COVID-19 hospital admissions data, SQUIREDL improved the accuracy of uncertainty-aware predictions by 22%, facilitating better resource planning. By focusing not only on model predictions but also on uncertainty estimates, SQUIREDL provided a more comprehensive understanding of potential patient outcomes, enabling more informed clinical decision-making. This approach also demonstrated the broad potential of computational healthcare applications.

6.1.2 Automating ML Workflows for Unlabelled Time Series

Research Question 2. *How can we develop automated learning frameworks that efficiently exploit unlabelled and partially labelled healthcare time series, while minimising human annotation effort and reducing dependence on expert supervision?*

Contribution 2. In Chapter 4 we developed SALTS, a novel approach to automate ML model training for biosignal time series, enhancing analysis efficiency. The motivation behind this study stemmed from the fact that labelling sensor-generated time series is challenging due to their sequential nature, requiring context and temporal dependencies to be annotated correctly. Unlike other domains, labelling biosignals requires medical expertise, which is costly and time-consuming. Current active learning strategies help, but they often treat annotation, hyperparameter tuning, and model training as separate processes, leading to inefficiency. Semi-supervised learning that uses unlabelled data with minimal human input can help, but it tends to limit accuracy as it fails to effectively incorporate the knowledge of domain experts. As such, automated workflows that integrate human expertise, unlabelled data, and model optimisation are lacking.

Recognising the costly nature of data labelling and that domain experts may have limited technical expertise in model optimisation, SALTS addresses these limitations by operating in a two-fold manner. It first operates at the data input level through adaptive

data acquisition, selecting highly informative samples for labelling. Then, it works at the model level, dynamically refining the model by exploring hyperparameter options and automatically maximising the use of unlabelled samples. This results in a robust learning strategy that continuously improves the model with expanding data and human expertise.

We demonstrated SALTS on EEG, ECG, and IMU time series classification, outperforming baselines and current state-of-the-art methods while reducing reliance on human input for model tuning. This enhances the applicability of machine learning to unlabelled or partially labelled healthcare time series, maximising the value of each human annotation step. This approach was successfully applied in both binary and multi-class classification tasks, showcasing its broad applicability to various healthcare use cases and ensuring improved results.

6.1.3 Ensuring Robustness under Distribution Shifts in Healthcare Time Series

Research Question 3. *How can we design models that generalise across heterogeneous time series distributions and remain effective under complex and overlapping distribution shifts in healthcare deployment settings?*

Contribution 3. In Chapter 5 we presented ADAPTOOD, a system for fine-tuning time series-based machine learning models for out-of-distribution (OOD) data in an uncertainty-aware manner. This was motivated by the fact that data samples collected in one context often differ from those encountered during deployment, and variations in hospitals, sensors, and patient populations cause distribution shifts that hinder model performance. Additionally, ECG datasets present additional complexities in labelling, making human-in-the-loop annotation solutions less effective. Pre-training methods help alleviate these issues, but their performance often degrades due to overfitting, especially when dealing with complex and multi-faceted distribution shifts. Effective solutions must recognise and quantify the severity of distribution shifts to improve model adaptation.

To address these challenges, ADAPTOOD leverages data uncertainty to quantify distribution shift severity, and uses this information to improve the fine-tuning process. To quantify the severity we leverage the differences between the pre-training and OOD data distributions. This form of uncertainty naturally correlates with the fine-grained OOD severity, making it a practical guidance signal for adaptation. Building on this, we use a combination of low-rank model updates and adaptive hyperparameter optimisation to further support the adaption mechanism.

ADAPTOOD effectively handles varying levels of distribution shifts, enabling robust, efficient, and accurate model updates across diverse OOD scenarios. In real-world ECG OOD use cases, the method delivered up to 7% higher accuracy and 12.9% higher precision compared to existing methods. ADAPTOOD outperformed alternatives and showed con-

sistent improvements in robustness, calibration, and generalisation, even in settings with varying scalability, complexity, or real-time deployment requirements. A key contributor to ADAPTOOD's success is its ability to adapt the degree of fine-tuning based on the severity of distribution shifts, highlighting the benefits of uncertainty-guided adaptation over static protocols. These results position ADAPTOOD as a promising and reliable approach for out-of-distribution model fine-tuning.

6.2 Key Insights and Broader Implications

The contributions examined above collectively address the research questions posed in Chapter 1 of this thesis, demonstrating the improvements made in machine learning solutions for sparse, unlabelled, and out-of-distribution biosignals. The findings and methods proposed can be used for the development and deployment of robust and generalisable machine learning systems that are applicable to numerous time series datasets. As such, the work presented in this thesis has promising implications for a wide set of stakeholders.

First, researchers can build upon the proposed methods and ideas to further develop ML systems with as little and as less-processed data as possible, from the healthcare domain and beyond. The versatility of the approaches examined can help in the development of scalable solutions, enriching the literature in the field. For example, medical professionals could use the suggested algorithms to build models for the easier diagnosis and prediction of pathologies both in clinical settings and in the wild. This is crucial for advancing healthcare in environments ranging from established medical centres to those in rural or underserved areas, as the ability to create effective models with minimal user intervention opens up new possibilities for physicians who may not have the specialised training to use traditional ML systems. Additionally, the algorithms can easily be adjusted to cater to specific populations, ensuring that the models are not only accurate but also equitable across demographics. With further development, individuals can use the outcomes and models whose development has been facilitated by this work to monitor their vital signs and overall health conditions continuously, improving their daily lives and making healthcare more accessible even in remote settings. This can empower users to stay informed about their health, facilitating early detection of potential medical issues and supporting healthier lifestyle choices. Therefore, as these technologies become more integrated into consumer devices such as wearables, they can seamlessly blend into everyday routines.

To continue, this thesis offers several implications for the advancement of ML applications in fields with real-world datasets, which are often irregular or unlabelled. This work has explored uncertainty-aware learning techniques that can effectively handle the challenges posed by real-world time series data. By incorporating uncertainty estimation, the models presented here have become more resilient to the challenging data commonly

encountered in healthcare scenarios, by not only improving the accuracy and reliability of predictions but also by enhancing the ability of ML models to adapt to varying data qualities. This makes the work better suited for deployment in diverse and uncontrolled environments, helping to bridge the gap between controlled research-based settings and the complex unpredictable nature of real-world monitoring.

While the methods developed in this thesis show strong promise, there are practical considerations that can potentially impact the deployment of any ML solution. For instance, the scale and heterogeneity of biosignals makes it challenging to scale ML models across diverse data types and populations. The ability to quantify uncertainty can guide model fine-tuning when new data becomes available, even in cases where there are underlying distribution shifts present, but future work must focus on how models can support long-term analysis and integrate more seamlessly into clinical and other real-world workflows. Through the development of strategies for efficiently leveraging sparse, unlabelled, and out-of-distribution data, this work has contributed to building ML pipelines that can be applied in real-world scenarios, and future research should ultimately focus on driving actionable interventions in a wider variety of settings. This work not only advances the state-of-the-art in the analysis of healthcare time series, but it also opens up new possibilities for similar challenges in other industries such as finance, environmental monitoring, and industrial applications. Overall, the approaches examined in this thesis can help in linking daily monitoring with decision-making processes, paving the way for more proactive, tailored, and timely responses.

6.3 Future Research Directions

As with any scientific endeavour, the findings presented in this thesis open up as many questions as they attempt to answer. While this thesis has made significant steps in contributing to the design of more generalisable, data-efficient, and adaptive algorithms for healthcare time series analysis, some intriguing avenues for future research remain. As such, there is a potential for future research to move beyond controlled experimental settings towards more dynamic use cases and environments, by integrating diverse sources of information, refining model interpretability, and grounding predictions in more realistic, human-centered contexts. Below we outline some key directions that could guide and inspire the next steps of this research.

- **Compatibility with Other Solutions.** The work presented in this thesis is focused on deep learning models designed around neural networks, and specifically on training those models from scratch or fine-tuning them as effectively as possible to explicitly address irregular sampling, missing annotations, and distribution shifts. While explicit ML model training is widely used at the time of writing this thesis, it

would be intriguing for future research to examine how to balance the strengths of specifically-trained models with the capabilities of Large Language Models (LLMs). These excel in tasks such as semantic reasoning, contextual comprehension, and content generation, but it would be valuable to explore how these capabilities could be leveraged to create a more informed backbone that still incorporates methods like those proposed. This could accommodate irregularly-sampled data and facilitate more accurate auto-labelling of unlabelled or semi-labelled datasets, while also incorporating the expertise of domain experts. In this context, rather than relying on the neural network-based models used in this thesis, future work might explore substituting them with LLMs to achieve the same objectives. This suggests that the future of LLMs may not lie in a single standalone strategy, but rather in hybrid approaches that combine the domain-specific strengths of specialised models with the generalisation capabilities of LLMs and other foundation models.

- **Generalisation to New Modalities.** While the contributions of Chapter 3 and Chapter 5 are time series-specific, it would be interesting to explore how the method proposed in Chapter 4 could be further developed to accommodate other data types as well, like imaging. The labelling difficulty of time series motivated Chapter 4, but given the potential that imaging approaches can have in AI-based healthcare applications like ophthalmological and radiological analysis, this would be an exciting future direction. Extending the framework beyond time series would enhance its generalisability to a wider extent, but also reveal how its uncertainty estimation and active learning components behave across different data structures. With appropriate adaptations, like using a ResNet model and adapting parts of the proposed architecture, SALTS could be applied to images or other modalities, to extend it and shift its focus more towards an AutoML-based training solution that makes the most of a user's annotation budget irrespective of the underlying data.

- **Sustainable Use of Resources.** Given the rate at which AI solutions are expanding, one of the most significant challenges in their real-world deployment is the availability of resources to train and deploy the necessary models. The solutions explored in this thesis do not lead to more energy consumption or carbon emissions than those associated with typical ML models, but as dataset sizes grow and tasks become more computationally intensive, their environmental footprint is an increasing concern. As such, a promising direction for future work would be to explore methods to optimise the energy efficiency of the training and fine-tuning processes described in this thesis while maintaining performance. The contribution of Chapter 5 employs strategies for reducing model size without compromising accuracy by limiting the number of trainable parameters, and future research could focus on expanding these strategies. This shift toward sustainability could also

extend to ML model reuse, through approaches for addressing distribution shifts like the one seen in Chapter 5. By leveraging larger pre-trained models and only fine-tuning them for specific tasks, rather than training them from scratch, energy consumption is minimised, as fine-tuning is possible with minimal computational resources. Ultimately, in the future the goal would be to combine the advances in irregular data handling and annotation approaches with practices that further prioritise sustainability, ensuring that AI-driven progress remains responsible.

6.4 Closing Remarks

This thesis demonstrates a significant advancement in the automation of the machine learning pipeline for time series data by investigating challenges related to uncertainty estimation, data labelling, and robustness under irregularly-sampled data or data under distribution shifts. The method proposed for incorporating uncertainty-aware learning into sequence-to-sequence predictions enhances the reliability and robustness of healthcare models when using real-world sensors data. Similarly, the approach explored for automating ML workflows and leveraging human expertise to efficiently label and refine models, offers an effective solution to the costly and time-consuming process of manual annotation while significantly improving performance. Further, the methodology developed for adapting models under distribution shifts ensures that predictive models can maintain their effectiveness across contexts and populations, thus enhancing their generalisability.

Overall, the works presented have shown that by better integrating sparse (irregularly-sampled) data, utilising unlabelled or limited labelled data, and increasing model resilience to distribution shifts, it is possible to develop more generalisable models for real-world time series applications. As the availability of time series data continues to grow and the use of machine learning becomes more widespread, the techniques developed in this thesis will play an even more significant role in creating accurate, robust, and effective AI-based models.

Bibliography

[1] C. Aggarwal, "Neural Networks and Deep Learning: A Textbook," *Springer International Publishing*, 2023.

[2] Y. Roh, G. Heo, and S. E. Whang, "A Survey on Data Collection for Machine Learning: A Big Data - AI Integration Perspective," *IEEE Transactions on Knowledge and Data Engineering*, vol. 33, no. 4, 2021.

[3] A. Mincholé, J. Camps, A. Lyon, and B. Rodríguez, "Machine Learning in the Electrocardiogram," *Journal of Electrocardiology*, vol. 57, pp. S61–S64, 2019.

[4] Y. Roy, H. Banville, I. Albuquerque, A. Gramfort, T. H. Falk, and J. Faubert, "Deep Learning-Based Electroencephalography Analysis: A Systematic Review," *Journal of Neural Engineering*, vol. 16, no. 5, p. 051001, Aug 2019.

[5] N. Gupta, S. K. Gupta, R. K. Pathak, V. Jain, P. Rashidi, and J. S. Suri, "Human Activity Recognition in Artificial Intelligence Framework: A Narrative Review," *Artificial Intelligence Review*, vol. 55, no. 6, p. 4755–4808, Jan 2022.

[6] D. Rodbard, "Continuous Glucose Monitoring: A Review of Successes, Challenges, and Opportunities," *Diabetes Technology & Therapeutics*, vol. 18, no. S2, 2016.

[7] S. N. Shukla and B. M. Marlin, "A Survey on Principles, Models and Methods for Learning from Irregularly Sampled Time Series," *NeurIPS Workshop on ML Retrospectives, Surveys & Meta-Analyses (ML-RSA)*, 2020.

[8] B. Settles, "Active Learning Literature Survey," University of Wisconsin–Madison, Computer Sciences Technical Report 1648, 2009.

[9] J. Yang, K. Zhou, Y. Li, and Z. Liu, "Generalized Out-of-Distribution Detection: A Survey," *International Journal of Computer Vision*, vol. 132, no. 12, p. 5635–5662, Jun 2024.

[10] J. Gawlikowski, C. R. N. Tassi, M. Ali, J. Lee, M. Humt, J. Feng, A. Kruspe, R. Triebel, P. Jung, R. Roscher, M. Shahzad, W. Yang, R. Bamler, and X. X. Zhu, "A Survey of Uncertainty in Deep Neural Networks," *Artificial Intelligence Review*, vol. 56, no. S1, p. 1513–1589, Jul 2023.

[11] L. J. L. López, S. Elsharief, D. A. Jorf, F. Darwish, C. Ma, and F. E. Shamout, "Uncertainty Quantification for Machine Learning in Healthcare: A Survey," *Conference on Health, Inference, and Learning (CHIL)*, 2025.

[12] E. v. E. Jesper and H. H. Hoos, "A Survey on Semi-Supervised Learning," *Springer Machine Learning*, vol. 109, p. 373–440, 2019.

[13] H. J. Escalante, "Automated Machine Learning: A Brief Review at the End of the Early Years," *Springer Automated Design of Machine Learning and Search Algorithms*, pp. 11–28, 2021.

[14] R. Dupre, J. Fajtl, V. Argyriou, and P. Remagnino, "Improving Dataset Volumes and Model Accuracy with Semi-Supervised Iterative Self-Learning," *IEEE Transactions on Image Processing*, vol. 29, pp. 4337–4348, 2020.

[15] C. Ding, T. Yao, C. Wu, and J. Ni, "Advances in Deep Learning for Personalized ECG Diagnostics: A Systematic Review Addressing Inter-Patient Variability and Generalization Constraints," *Biosensors and Bioelectronics*, vol. 271, p. 117073, 2025.

[16] U. Gupta, N. Paluru, D. Nankani, K. Kulkarni, and N. Awasthi, "A Comprehensive Review on Efficient Artificial Intelligence Models for Classification of Abnormal Cardiac Rhythms using Electrocardiograms," *Heliyon*, vol. 10, no. 5, p. e26787, 2024.

[17] M. Gholizade, H. Soltanizadeh, M. Rahmanimanesh, and S. S. Sana, "A Review of Recent Advances and Strategies in Transfer Learning," *International Journal of System Assurance Engineering and Management*, vol. 16, no. 3, p. 1123–1162, Feb 2025.

[18] A. Hosna, E. Merry, J. Gyalmo, Z. Alom, Z. Aung, and M. A. Azim, "Transfer Learning: A Friendly Introduction," *Journal of Big Data*, vol. 9, no. 1, Oct 2022.

[19] A. Farahani, S. Voghoei, K. Rasheed, and H. R. Arabnia, "A Brief Review of Domain Adaptation," *Advances in Data Science and Information Engineering*, pp. 877–894, 2021.

[20] S. Rajendran, W. Pan, M. R. Sabuncu, Y. Chen, J. Zhou, and F. Wang, "Learning Across Diverse Biomedical Data Modalities and Cohorts: Challenges and Opportunities for Innovation," *Patterns*, vol. 5, no. 2, p. 100913, Feb 2024.

[21] S. Vavaroutas, T. Dang, E. Rocheteau, and C. Mascolo, "SQUIREDL: Sparse Sequence-to-Sequence Uncertainty Estimation in Evidential Deep Learning," *ACM Transactions on Computing for Healthcare*, vol. 6, no. 3, 2025.

[22] S. Vavaroutas, G. Rizos, and C. Mascolo, "SALTS: Streamlined Adaptive Learning for Sensors Time Series," in *Proceedings of the 47th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2025.

[23] S. Vavaroutas, Y. Wu, A. Etemad, and C. Mascolo, "ADAPTOOD: Uncertainty-Aware Fine-Tuning for Out-of-Distribution ECG Time Series Models," 2026.

[24] S. Vavaroutas, L. Qendro, and C. Mascolo, "Uncertainty Estimation with Data Augmentation for Active Learning Tasks on Health Data," in *Proceedings of the 45th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2023.

[25] S. Vavaroutas, T. Dang, E. Rocheteau, and C. Mascolo, "Uncertainty Estimation for Sequence-to-Sequence Regression on Sparse Time Series," *5th UK Mobile, Wearable and Ubiquitous Systems Research Symposium (MobiUK)*, 2023.

[26] A. Alvarez Olmo, S. Vavaroutas, Y. Wu, and C. Mascolo, "Benchmarking Foundation Models on Out-of-Distribution Wearable Biosignals," *7th UK Mobile, Wearable and Ubiquitous Systems Research Symposium (MobiUK)*, 2025.

[27] Y. J. Lee, C. Park, H. Kim, S. J. Cho, and W.-H. Yeo, "Artificial Intelligence on Biomedical Signals: Technologies, Applications, and Future Firections," *Med-X*, vol. 2, no. 1, Dec 2024.

[28] X.-L. Yang, G.-Z. Liu, Y.-H. Tong, H. Yan, Z. Xu, Q. Chen, X. Liu, H.-H. Zhang, H.-B. Wang, and S.-H. Tan, "The History, Hotspots, and Trends of Electrocardiogram," *Journal of Geriatric Cardiology*, vol. 12, no. 4, pp. 448–456, 2015.

[29] A. Arjoonsingh, B. C. Jamal, and L. Ganti, "History and Evolution of the Electroencephalogram," *Cureus*, Aug 2024.

[30] D. Ma, A. Ferlini, and C. Mascolo, "Innovative Human Motion Sensing With Earbuds," *GetMobile: Mobile Computing and Communications*, vol. 25, no. 4, Mar 2022.

[31] A. Ferlini, D. Ma, L. Qendro, and C. Mascolo, "Mobile Health With Head-Worn Devices: Challenges and Opportunities," *IEEE Pervasive Computing*, vol. 21, no. 3, 2022.

[32] P. Picerno, M. Iosa, C. D'Souza, M. G. Benedetti, S. Paolucci, and G. Morone, "Wearable Inertial Sensors for Human Movement Analysis: A Five-Year Update," *Expert Review of Medical Devices*, vol. 18, no. sup1, pp. 79–94, 2021.

[33] C. Bender, P. Vestergaard, and S. L. Cichosz, "The History, Evolution and Future of Continuous Glucose Monitoring (CGM)," *Diabetology*, vol. 6, no. 3, 2025.

[34] A. Madan, M. Cebrian, S. Moturu, K. Farrahi, and A. S. Pentland, "Sensing the "Health State" of a Community," *IEEE Pervasive Computing*, vol. 11, no. 4, 2012.

[35] K.-J. Butkow, T. Dang, A. Ferlini, D. Ma, Y. Liu, and C. Mascolo, "An Evaluation of Heart Rate Monitoring with In-Ear Microphones Under Motion," *Pervasive and Mobile Computing*, vol. 100, p. 101913, May 2024.

[36] K. Farrahi and D. Gatica-Perez, "What Did You Do Today? Discovering Daily Routines from Large-Scale Mobile Data," in *Proceedings of the 16th ACM International Conference on Multimedia*. Association for Computing Machinery, 2008.

[37] I. Jeong, W. G. Chung, E. Kim, W. Park, H. Song, J. Lee, M. Oh, E. Kim, J. Paek, T. Lee, D. Kim, S. H. An, S. Kim, H. Cho, and J.-U. Park, "Machine Learning in Biosignal Analysis from Wearable Devices," *Materials Horizons*, vol. 12, no. 17, p. 6587–6621, 2025.

[38] R. Alsagri, A. Wilde, K. Farrahi, A. Alhomoud, and N. White, "Emerging Technologies and Trends in Lung Volume Estimation: A Review," *IEEE Sensors Journal*, vol. 25, no. 13, 2025.

[39] J. Schmidhuber, "Deep Learning in Neural Networks: An Overview," *Neural Networks*, vol. 61, pp. 85–117, 2015.

[40] H. Ismail Fawaz, G. Forestier, J. Weber, L. Idoumghar, and P.-A. Muller, "Deep Learning for Time Series Classification: A Review," *Data Mining and Knowledge Discovery*, vol. 33, no. 4, p. 917–963, Mar 2019.

[41] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning Representations by Back-Propagating Errors," *Nature*, vol. 323, no. 6088, p. 533–536, Oct 1986.

[42] R. Pascanu, T. Mikolov, and Y. Bengio, "On the Difficulty of Training Recurrent Neural Networks," in *Proceedings of the 30th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 28, no. 3. PMLR, 2013, pp. 1310–1318.

[43] K. O'Shea and R. Nash, "An Introduction to Convolutional Neural Networks," *arXiv*, vol. 1511.08458, Dec 2015.

[44] M. Atencia, R. Stoean, and G. Joya, "Uncertainty Quantification through Dropout in Time Series Prediction by Echo State Networks," *Mathematics*, vol. 8, no. 8, Jun 2020.

[45] U. Kummaraka and P. Srisuradetchai, "Monte Carlo Dropout Neural Networks for Forecasting Sinusoidal Time Series: Performance Evaluation and Uncertainty Quantification," *Applied Sciences*, vol. 15, no. 8, 2025.

[46] S. Yadav and V. Subbian, "Monte Carlo ExtremalMask: Uncertainty Aware Time Series Model Interpretability For Critical Care Applications," in *Proceedings of the 10th Machine Learning for Healthcare Conference*, ser. Proceedings of Machine Learning Research, vol. 298, 2025.

[47] D. Wu, L. Gao, M. Chinazzi, X. Xiong, A. Vespignani, Y.-A. Ma, and R. Yu, "Quantifying Uncertainty in Deep Spatiotemporal Forecasting," in *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. Association for Computing Machinery, 2021, p. 1841–1851.

[48] M. Magris and A. Iosifidis, "Bayesian Learning for Neural Networks: An Algorithmic Survey," *Artificial Intelligence Review*, vol. 56, no. 10, p. 11773–11823, 2023.

[49] Q. Pan, P. Yang, and J. Zhang, "BayesTSF: Measuring Uncertainty Estimation in Industrial Time Series Forecasting from a Bayesian Perspective," in *Advanced Intelligent Computing Technology and Applications*. Springer Nature Singapore, 2024, pp. 81–93.

[50] K. Wickstrom, K. O. Mikalsen, M. Kampffmeyer, A. Revhaug, and R. Jenssen, "Uncertainty-Aware Deep Ensembles for Reliable and Explainable Predictions of Clinical Time Series," *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 7, pp. 2435–2444, Dec 2020.

[51] Z. Abulawi, R. Hu, P. Balaprakash, and Y. Liu, "Bayesian Optimized Deep Ensemble for Uncertainty Quantification of Deep Neural Networks: a System Safety Case Study on Sodium Fast Reactor Thermal Stratification Modeling," *Reliability Engineering & System Safety*, vol. 264, p. 111353, 2025.

[52] M. Fontana, G. Zeni, and S. Vantini, "Conformal Prediction: A Unified Review of Theory and New Challenges," *Bernoulli*, vol. 29, no. 1, pp. 1 – 23, 2023.

[53] X. Zhou, B. Chen, Y. Gui, and L. Cheng, "Conformal Prediction: A Data Perspective," *ACM Comput. Surv.*, vol. 58, no. 2, Sep 2025.

[54] J. Vazquez and J. C. Facelli, "Conformal Prediction in Clinical Medical Sciences," *Journal of Healthcare Informatics Research*, vol. 6, no. 3, p. 241–252, 2022.

[55] E. Straitouri, L. Wang, N. Okati, and M. Gomez Rodriguez, "Improving Expert Predictions with Conformal Prediction," in *Proceedings of the 40th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 202. PMLR, 2023, pp. 32 633–32 653.

[56] A. Portela, J. R. Banga, and M. Matabuena, "Conformal Prediction for Uncertainty Quantification in Dynamic Biological Systems," *PLOS Computational Biology*, vol. 21, no. 5, p. e1013098, 2025.

[57] M. Zaffran, O. Feron, Y. Goude, J. Josse, and A. Dieuleveut, "Adaptive Conformal Predictions for Time Series," in *Proceedings of the 39th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 162. PMLR, 2022, pp. 25 834–25 866.

[58] A. Amini, W. Schwarting, A. Soleimany, and D. Rus, "Deep Evidential Regression," in *Advances in Neural Information Processing Systems*, vol. 33. Curran Associates, Inc., 2020, pp. 14 927–14 937.

[59] M. Sensoy, L. Kaplan, and M. Kandemir, "Evidential Deep Learning to Quantify Classification Uncertainty," in *Advances in Neural Information Processing Systems*, vol. 31. Curran Associates, Inc., 2018.

[60] T. Xia, J. Han, L. Qendro, T. Dang, and C. Mascolo, "Hybrid-EDL: Improving Evidential Deep Learning for Uncertainty Quantification on Imbalanced Data," *NeurIPS Workshop on Trustworthy and Socially Responsible ML*, 2022.

[61] R. T. Q. Chen, Y. Rubanova, J. Bettencourt, and D. K. Duvenaud, "Neural Ordinary Differential Equations," in *Advances in Neural Information Processing Systems*, vol. 31. Curran Associates, Inc., 2018.

[62] Y. Rubanova, R. T. Chen, and D. K. Duvenaud, "Latent Ordinary Differential Equations for Irregularly-Sampled Time Series," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 32, 2019.

[63] P. Kidger, J. Morrill, J. Foster, and T. Lyons, "Neural Controlled Differential Equations for Irregular Time Series," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, pp. 6696–6707, 2020.

[64] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, Nov 1997.

[65] A. Alaa and M. Van Der Schaar, "Frequentist Uncertainty in Recurrent Neural Networks via Blockwise Influence Functions," in *Proceedings of the 37th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 119, Jul 2020, pp. 175–190.

[66] D. D. Lewis and J. Catlett, "Heterogeneous Uncertainty Sampling for Supervised Learning," *ML Proceedings '94*, pp. 148–156, 1994.

[67] Y. Gal, R. Islam, and Z. Ghahramani, "Deep Bayesian Active Learning with Image Data," in *Proceedings of the 34th International Conference on Machine Learning (ICML)*, vol. 70, 2017, pp. 1183–1192.

[68] Y. Ovadia, E. Fertig, J. Ren, Z. Nado, D. Sculley, S. Nowozin, J. Dillon, B. Lakshminarayanan, and J. Snoek, "Can You Trust Your Model's Uncertainty? Evaluating Predictive Uncertainty Under Dataset Shift," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 32, 2019.

[69] A. de Mathelin, F. Deheeger, M. Mougeot, and N. Vayatis, "Deep Out-of-Distribution Uncertainty Quantification via Weight Entropy Maximization," *Journal of Machine Learning Research*, vol. 26, no. 4, pp. 1–68, 2025.

[70] S. H. Gheshlaghi, N. Y. Soltani, and M. Ganji, "Uncertainty Estimation for Out-of-Distribution Detection of Whole Slide Images," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2025)*, 2025, pp. 1–5.

[71] M. S. Tackney, D. G. Cook, D. Stahl, K. Ismail, E. Williamson, and J. Carpenter, "Framework for Handling Missing Accelerometer Outcome Data," *Trials*, vol. 22, no. 1, 2021.

[72] L. E. Alejo-Sanchez, A. Márquez-Grajales, F. Salas-Martínez, A. Franco-Arcega, V. López-Morales, O. A. Acevedo-Sandoval, C. A. González-Ramírez, and R. Villegas-Vega, "Missing Data Imputation of Climate Time Series: A Review," *MethodsX*, vol. 15, p. 103455, 2025.

[73] H. Akima, "A Method of Bivariate Interpolation and Smooth Surface Fitting for Irregularly Distributed Data Points," *ACM Trans. Math. Softw.*, vol. 4, no. 2, p. 148–159, Jun 1978.

[74] The SciPy Community, "Akima1DInterpolator," *SciPy Docs*, 2019.

[75] S. Lin, W. Lin, W. Wu, H. Chen, and C. L. P. Chen, "SparseTSF: Lightweight and Robust Time Series Forecasting via Sparse Modeling," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 48, no. 1, pp. 170–183, 2026.

[76] R. Peterson and J. Cavanaugh, "Fast, Effective, and Coherent Time Series Modelling using the Sparsity-Ranked Lasso," *Statistical Modelling*, vol. 25, no. 2, p. 150–169, 2025.

[77] J. Xu, K. Wang, C. Lin, L. Xiao, X. Huang, and Y. Zhang, "FM-GRU: A Time Series Prediction Method for Water Quality Based on seq2seq Framework," *Water*, vol. 13, no. 8, 2021.

[78] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to Sequence Learning with Neural Networks," in *Advances in Neural Information Processing Systems*, vol. 27. Curran Associates, Inc., 2014.

[79] J. Chorowski and N. Jaitly, "Towards Better Decoding and Language Model Integration in Sequence to Sequence Models," in *Interspeech 2017*. ISCA, 2017, p. 523–527.

[80] C. Xiao, X. Jiang, X. Du, W. Yang, W. Lu, X. Wang, and K. Chetty, "Boundary-Enhanced Time Series Data Imputation with Long-Term Dependency Diffusion Models," *Knowledge-Based Systems*, vol. 310, p. 112917, 2025.

[81] S. Xu, Y. Wang, X. Xu, G. Shi, H. Huang, and Y. Zheng, "A PatchTST-GRU Based Heterogeneous Seq2Seq Model with Numerical Weather Prediction Refinement for Multi-Step Wind Power Forecasting," *Scientific Reports*, vol. 15, no. 1, 2025.

[82] W. K. Wang, I. Chen, L. Hershkovich, J. Yang, A. Shetty, G. Singh, Y. Jiang, A. Kotla, J. Z. Shang, R. Yerrabelli, A. R. Roghanizad, M. M. H. Shandhi, and J. Dunn, "A Systematic Review of Time Series Classification Techniques Used in Biomedical Applications," *Sensors*, vol. 22, no. 20, 2022.

[83] Q. Tan, M. Ye, A. J. Ma, B. Yang, T. C.-F. Yip, G. L.-H. Wong, and P. C. Yuen, "Explainable Uncertainty-Aware Convolutional Recurrent Neural Network for Irregular Medical Time Series," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 10, pp. 4665–4679, Oct 2021.

[84] E. Rocheteau, P. Liò, and S. Hyland, "Temporal Pointwise Convolutional Networks for Length of Stay Prediction in the Intensive Care Unit," in *Proceedings of the Conference on Health, Inference, and Learning*, 2021, p. 58–68.

[85] Z. Che, S. Purushotham, K. Cho, D. Sontag, and Y. Liu, "Recurrent Neural Networks for Multivariate Time Series with Missing Values," *Scientific Reports*, vol. 8, no. 1, 2018.

[86] R. Miotto, F. Wang, S. Wang, X. Jiang, and J. T. Dudley, "Deep Learning for Healthcare: Review, Opportunities and Challenges," *Briefings in Bioinformatics*, vol. 19, no. 6, pp. 1236–1246, May 2017.

[87] M.-A. Zöller and M. F. Huber, "Benchmark and Survey of Automated Machine Learning Frameworks," *Journal of Artificial Intelligence Research*, vol. 70, p. 409–472, Jan 2021.

[88] A. Esteva, A. Robicquet, B. Ramsundar, V. Kuleshov, M. DePristo, K. Chou, C. Cui, G. Corrado, S. Thrun, and J. Dean, "A Guide to Deep Learning in Healthcare," *Nature Medicine*, vol. 25, no. 1, p. 24–29, Jan 2019.

[89] E. J. Topol, "High-Performance Medicine: The Convergence of Human and Artificial Intelligence," *Nature Medicine*, vol. 25, no. 1, p. 44–56, Jan 2019.

[90] F. Peng, Q. Luo, and L. M. Ni, "ACTS: An Active Learning Method for Time Series Classification," *IEEE 33rd International Conference on Data Engineering (ICDE)*, pp. 175–178, 2017.

[91] R. Adaimi and E. Thomaz, "Leveraging Active Learning and Conditional Mutual Information to Minimize Data Annotation in Human Activity Recognition," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT)*, vol. 3, no. 3, 2019.

[92] C. A. Flores and R. Verschae, "A Generic Semi-Supervised and Active Learning Framework for Biomedical Text Classification," in *Proceedings of the 44th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2022.

[93] S. Roy, R. Mehera, R. K. Pal, and S. K. Bandyopadhyay, "Hyperparameter Optimization for Deep Neural Network Models: A Comprehensive Study on Methods and Techniques," *Innovations in Systems and Software Engineering*, no. 3, p. 789–800, Sep 2025.

[94] J. Snoek, H. Larochelle, and R. P. Adams, "Practical Bayesian Optimization of Machine Learning Algorithms," in *Advances in Neural Information Processing Systems*, vol. 25. Curran Associates, Inc., 2012.

[95] L. Franceschi, M. Donini, V. Perrone, A. Klein, C. Archambeau, M. Seeger, M. Pontil, and P. Frasconi, "Hyperparameter Optimization in Machine Learning," *Foundations and Trends in Machine Learning*, vol. 18, no. 6, pp. 975–1109, 10 2025.

[96] M. A. K. Raiaan, S. Sakib, N. M. Fahad, A. A. Mamun, M. A. Rahman, S. Shatabda, and M. S. H. Mukta, "A Systematic Review of Hyperparameter Optimization Techniques in Convolutional Neural Networks," *Decision Analytics Journal*, vol. 11, p. 100470, 2024.

[97] L. Yang and A. Shami, "On Hyperparameter Optimization of Machine Learning Algorithms: Theory and Practice," *Neurocomputing*, vol. 415, pp. 295–316, Nov 2020.

[98] A. Alsharef, K. Aggarwal, Sonia, M. Kumar, and A. Mishra, "Review of ML and AutoML Solutions to Forecast Time-Series Data," *Archives of Computational Methods in Engineering*, vol. 29, no. 7, 2022.

[99] J. Waring, C. Lindvall, and R. Umeton, "Automated Machine Learning: Review of the State-of-the-Art and Opportunities for Healthcare," *Artificial Intelligence in Medicine*, vol. 104, p. 101822, 2020.

[100] A. Arefeen and H. Ghasemzadeh, "Cost-Effective Multitask Active Learning in Wearable Sensor Systems," *Sensors*, no. 5, 2025.

[101] K. Bhardwaj, N. Goyal, B. Mittal, V. Sharma, and S. N. Shivhare, "A Novel Active Learning Technique for Fetal Health Classification Based on XGBoost Classifier," *IEEE Access*, vol. 13, pp. 9485–9497, 2025.

[102] M. Santos and G. Marreiros, "A Systematic Review of Active Learning Approaches in the Selection of Medical Images," *Procedia Computer Science*, vol. 256, pp. 843–851, 2025.

[103] X. Chen and B. Wujek, "A Unified Framework for Automatic Distributed Active Learning," *IEEE Trans. on Pattern Analysis*, 2022.

[104] P. Koch, O. Golovidov, S. Gardner, B. Wujek, J. Griffin, and Y. Xu, "Autotune: A Derivative-Free Optimization Framework for Hyperparameter Tuning," *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, p. 443–452, 2018.

[105] J. D. Griffin and T. G. Kolda, "Nonlinearly-Constrained Optimization using Heuristic Penalty Methods and Asynchronous Parallel Generating Set Search," *Applied Mathematics Research*, pp. 36–62, 2010.

[106] C. Yang, Y. Akimoto, D. W. Kim, and M. Udell, "OBOE: Collaborative Filtering for AutoML Model Selection," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, ser. KDD '19. Association for Computing Machinery, 2019, p. 1173–1183.

[107] H. Amiri, J. Mohammadzadeh, S. Mohsen Mirhosseini, and A. Nikravanshalmani, "Prediction of High-Risk Cardiac Arrhythmia Based on Optimized Deep Active Learning," *IEEE Access*, vol. 13, pp. 39 006–39 032, 2025.

[108] D. Holtz, C. Kaymakci, D. Leuthe, S. Wenninger, and A. Sauer, "A Data-Efficient Active Learning Architecture for Anomaly Detection in Industrial Time Series Data," *Flexible Services and Manufacturing Journal*, Feb 2025.

[109] C. I. Tang, I. Perez-Pozuelo, D. Spathis, S. Brage, N. Wareham, and C. Mascolo, "SelfHAR: Improving Human Activity Recognition through Self-training with Unlabeled Data," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2021.

[110] A. Singh, A. J. Rana, A. Kumar, S. Vyas, and Y. S. Rawat, "Semi-Supervised Active Learning for Video Action Detection," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 5, p. 4891–4899, Mar 2024.

[111] B. Jiao, H. M. Gomes, B. Xue, Y. Guo, and M. Zhang, "A Semi-Supervised Active Learning Neural Network for Data Streams With Concept Drift," *IEEE Computational Intelligence Magazine*, vol. 20, no. 1, pp. 18–33, 2025.

[112] J. Lim, J. Na, and N. Kwak, "Active Semi-Supervised Learning by Exploring Per-Sample Uncertainty and Consistency," *arXiv*, vol. 2303.08978, 2023.

[113] M. Gao, Z. Zhang, G. Yu, S. Ö. Arik, L. S. Davis, and T. Pfister, "Consistency-Based Semi-Supervised Active Learning: Towards Minimizing Labeling Cost," *European Conference on Computer Vision (ECCV)*, pp. 510–526, 2020.

[114] L.-Z. Guo, L.-H. Jia, J.-J. Shao, and Y.-F. Li, "Robust Semi-Supervised Learning in Open Environments," *Frontiers of Computer Science*, vol. 19, no. 8, Jan 2025.

[115] P. Yue, Z. Li, M. Zhou, X. Wang, and P. Yang, "Wearable-Sensor-Based Weakly Supervised Parkinson's Disease Assessment with Data Augmentation," *Sensors*, vol. 24, no. 4, p. 1196, 2024.

[116] A. Ratner, S. H. Bach, H. Ehrenberg, J. Fries, S. Wu, and C. Ré, "Snorkel: Rapid Training Data Creation with Weak Supervision," *Proceedings of the VLDB Endowment*, vol. 11, no. 3, p. 269–282, 2017.

[117] I. Harmon, "A Brief Overview of Weak Supervision," *Data Science Research*, 2020.

[118] M. Cauchois, S. Gupta, A. Ali, and J. C. Duchi, "Predictive Inference with Weak Supervision," *Journal of Machine Learning Research*, vol. 25, no. 118, pp. 1–45, 2024.

[119] E. J. da S. Luz, W. R. Schwartz, G. Cámara-Chávez, and D. Menotti, "ECG-Based Heartbeat Classification for Arrhythmia Detection: A Survey," *Computer Methods and Programs in Biomedicine*, vol. 127, pp. 144–164, 2016.

[120] T. Nakamura and T. Sasano, "Emerging Potential and Challenges of AI-Based ECG Analysis in Clinical Medicine," *JACC Asia*, vol. 5, no. 1, Part 1, pp. 99–100, 2025.

[121] J. Wu and J. He, "Trustworthy Transfer Learning: Transferability and Trustworthiness," in *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. Association for Computing Machinery, 2023, p. 5829–5830.

[122] L. Xu, M. Xu, Y. Ke, X. An, S. Liu, and D. Ming, "Cross-Dataset Variability Problem in EEG Decoding with Deep Learning," *Frontiers in Human Neuroscience*, vol. 14, Apr 2020.

[123] W. M. Kouw and M. Loog, "An Introduction to Domain Adaptation and Transfer Learning," *arXiv*, vol. 1812.11806, Jan 2019.

[124] J. Ai and Z. Ren, "Not All Distributional Ahifts are Equal: Fine-Grained Robust Conformal Inference," in *Proceedings of the 41st International Conference on Machine Learning (ICML)*, vol. 235. PMLR, Jul 2024, pp. 641–665.

[125] P. S. Stumpf, X. Du, H. Imanishi, Y. Kunisaki, Y. Semba, T. Noble, R. C. G. Smith, M. Rose-Zerili, J. J. West, R. O. C. Oreffo, K. Farrahi, M. Niranjan, K. Akashi, F. Arai, and B. D. MacArthur, "Transfer Learning Efficiently Maps Bone Marrow Cell Types from Mouse to Human using Single-Cell RNA Sequencing," *Communications Biology*, vol. 3, no. 1, Dec 2020.

[126] A. Althnian, D. AlSaeed, H. Al-Baity, A. Samha, A. B. Dris, N. Alzakari, A. Abou Elwafa, and H. Kurdi, "Impact of Dataset Size on Classification Performance: An Empirical Evaluation in the Medical Domain," *Applied Sciences*, vol. 11, no. 2, Jan 2021.

[127] C. Li, Y. Liu, T. Denison, and T. Zhu, "BioX-Bridge: Model Bridging for Unsupervised Cross-Modal Knowledge Transfer Across Biosignals," *arXiv*, vol. 2510.02276, Oct 2025.

[128] F. Wenzel, A. Dittadi, P. Gehler, C.-J. Simon-Gabriel, M. Horn, D. Zietlow, D. Kernert, C. Russell, T. Brox, B. Schiele, B. Schölkopf, and F. Locatello, "Assaying Out-of-Distribution Generalization in Transfer Learning," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 35, pp. 7181–7198, 2022.

[129] M. Deng, V. Kyrki, and D. Baumann, "Transfer Learning in Latent Contextual Bandits with Covariate Shift through Causal Transportability," in *Proceedings of the Fourth Conference on Causal Learning and Reasoning*, ser. Proceedings of Machine Learning Research, vol. 275. PMLR, 2025, pp. 731–756.

[130] H. Sun, Z. Xie, H.-Y. He, and M. Li, "Mitigating Negative Transfer via Reducing Environmental Disagreement," *arXiv*, vol. 2510.24044, Oct 2025.

[131] H. M. Pham, A. Saeed, and D. Ma, "Revisiting Masked Auto-Encoders for ECG-Language Representation Learning," *The First NeurIPS Workshop on Time Series in the Age of Large Models*, vol. 1806.07366, Dec 2024.

[132] Y. Bengio, A. C. Courville, and P. Vincent, "Representation Learning: A Review and New Perspectives," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, pp. 1798–1828, 2012.

[133] W. Lu, J. Wang, X. Sun, Y. Chen, and X. Xie, "Out-of-Distribution Representation Learning for Time Series Classification," *International Conference on Learning Representations (ICLR)*, 2023.

[134] T. Mehari and N. Strodthoff, "Self-Supervised Representation Learning from 12-Lead ECG Data," *Computers in Biology and Medicine*, vol. 141, p. 105114, 2022.

[135] P. Trirat, Y. Shin, J. Kang, Y. Nam, J. Na, M. Bae, J. Kim, B. Kim, and J.-G. Lee, "Universal Time-Series Representation Learning: A Survey," *arXiv*, vol. 2401.03717, Aug 2024.

[136] J.-P. Jiang, S.-Y. Liu, H.-R. Cai, Q. Zhou, and H.-J. Ye, "Representation Learning for Tabular Data: A Comprehensive Survey," *arXiv*, vol. 2504.16109, 2025.

[137] Y. Qin, X. Zhang, S. Yu, and G. Feng, "A Survey on Representation Learning for Multi-View Data," *Neural Networks*, vol. 181, p. 106842, 2025.

[138] B. Chidlovskii, S. Clinchant, and G. Csurka, "Domain Adaptation in the Absence of Source Domain Data," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*. Association for Computing Machinery, 2016, p. 451–460.

[139] H. He, O. Queen, T. Koker, C. Cuevas, T. Tsiligkaridis, and M. Zitnik, "Domain Adaptation for Time Series under Feature and Label Shifts," in *Proceedings of the 40th International Conference on Machine Learning (ICML)*, vol. 202, Jul 2023, pp. 12 746–12 774.

[140] Y.-C. Yu and H.-T. Lin, "Semi-Supervised Domain Adaptation with Source Label Adaptation," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 24 100–24 109.

[141] W. Fan, Y. Si, M. Sun, L. Zhou, W. Yang, A. Alhudhaif, and F. Alenezi, "Class-Specific Weighted Broad Learning System-Based Domain Adaptation for Patient-Specific ECG Classification," *Expert Systems with Applications*, vol. 273, p. 126824, 2025.

[142] R. An, Z. Shang, Y. Liu, X. Wu, J. Ren, and Z. Zhao, "Imbalanced Domain Adaptation Net for Multi-Class Electrocardiography Signals Classification," *Biomedical Signal Processing and Control*, vol. 108, p. 107912, 2025.

[143] Y. Lu, H. Huang, X. Hu, and Z. Lai, "Multiple Adaptation Network for Multi-Source and Multi-Target Domain Adaptation," *IEEE Transactions on Multimedia*, vol. 27, pp. 5731–5745, 2025.

[144] C. Choi, Y. Lee, A. Chen, A. Zhou, A. Raghunathan, and C. Finn, "AutoFT: Learning an Objective for Robust Fine-Tuning," *NeurIPS Workshop on Distribution Shifts*, 2024.

[145] E. Svensson, H. R. Friesacher, A. Arany, L. Mervin, and O. Engkvist, "Temporal Evaluation of Uncertainty Quantification Under Distribution Shift," *AI in Drug Discovery*, pp. 132–148, 2025.

[146] V. Kuleshov, N. Fenner, and S. Ermon, "Accurate Uncertainties for Deep Learning using Calibrated Regression," in *PMLR*, vol. 80, 2018, pp. 2796–2804.

[147] G. Shafiee, M. Mohajeri-Tehrani, M. Pajouhi, and B. Larijani, "The Importance of Hypoglycemia in Diabetic Patients," *Diabetes and Metabolic Disorders*, vol. 11, no. 1, 2012.

[148] D. Dermawan and P. M. A. Kenichi, "An Overview of Advancements in Closed-Loop Artificial Pancreas System," *Heliyon*, vol. 8, no. 11, Nov 2022.

[149] M. Burke, "Life Expectancy for Type 1 Diabetes," *EndocrineWeb*, 2022.

[150] M. K. Danesh, E. Garosi, and H. Golmohamadpour, "The COVID-19 Pandemic and Nursing Challenges: A Review of the Early Literature," *Work*, vol. 69, no. 1, p. 23–36, 2021.

[151] A. Gordon and A. Reingold, "The Burden of Influenza: A Complex Problem," *Current Epidemiology Reports*, vol. 5, no. 1, pp. 1–9, Feb 2018.

[152] D. G. T. Denison, C. C. Holmes, B. K. Mallick, and A. F. M. Smith, "Bayesian Methods for Nonlinear Classification and Regression," *John Wiley & Sons*, Mar 2002.

[153] M. Minderer, J. Djolonga, R. Romijnders, F. Hubis, X. Zhai, N. Houlsby, D. Tran, and M. Lucic, "Revisiting the Calibration of Modern Neural Networks," in *Advances in Neural Information Processing Systems*, vol. 34, 2021, pp. 15 682–15 694.

[154] K. Rufibach, "Use of Brier score to Assess Binary Predictions," *Journal of Clinical Epidemiology*, vol. 63, no. 8, pp. 938–939, 2010.

[155] J. Nixon, M. W. Dusenberry, L. Zhang, G. Jerfel, and D. Tran, "Measuring Calibration in Deep Learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019.

[156] SKLearn Developers, "API Reference: Paired Euclidean Distances," *Scikit-learn*, 2007.

[157] M. Ott, M. Auli, D. Grangier, and M. Ranzato, "Analyzing Uncertainty in Neural Machine Translation," in *Proceedings of the 35th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 80. PMLR, 2018, pp. 3956–3965.

[158] P. A. Bosman and D. Thierens, "Negative Log-Likelihood and Statistical Hypothesis Testing as the Basis of Model Selection in IDEAs," *Technical Report*, Aug 2000.

[159] W. J. Maddox, P. Izmailov, T. Garipov, D. P. Vetrov, and A. G. Wilson, "A Simple Baseline for Bayesian Uncertainty in Deep Learning," in *Advances in Neural Information Processing Systems*, vol. 32. Curran Associates, Inc., 2019.

[160] SKLearn Developers, "API Reference: Explained Variance Score," *Scikit-learn*, 2010.

[161] L. Erlygin, V. Zholobov, V. Baklanova, E. Sokolovskiy, and A. Zaytsev, "Surrogate Uncertainty Estimation for your Time Series Forecasting Black-Box: Learn When to Trust," in *2023 IEEE International Conference on Data Mining Workshops (ICDMW)*, 2023, pp. 1247–1258.

[162] Z. Xiao, "Time Series Quantile Regressions," in *Time Series Analysis: Methods and Applications*, ser. Handbook of Statistics. Elsevier, 2012, vol. 30, pp. 213–257.

[163] SKLearn Developers, "API Reference: SKLearn Metrics," *Scikit-learn*, 2007.

[164] N. Bacaër, "Lotka, Volterra and the Predator-Prey System," *A Short History of Mathematical Population Dynamics*, pp. 71–76, 2011.

[165] E. Mahase, "Type 1 Diabetes: Global Prevalence is Set to Double by 2040, Study Estimates," *BMJ*, p. o2289, Sep 2022.

[166] L. Crenier, C. Abou-Elias, and B. Corvilain, "Glucose Variability Assessed by Low Blood Glucose Index is Predictive of Hypoglycemic Events in Patients with Type 1 Diabetes Switched to Pump Therapy," *Diabetes Care*, vol. 36, no. 8, pp. 2148–2153, Jul 2013.

[167] C. D. Man, F. Micheletto, D. Lv, M. Breton, B. Kovatchev, and C. Cobelli, "UVA/PADOVA Diabetes Simulator," *Diabetes and Technology*, vol. 8, no. 1, pp. 26–34, 2014.

[168] R. Visentin, C. D. Man, B. Kovatchev, and C. Cobelli, "The University of Virginia/PADOVA Type 1 Diabetes Simulator Matches the Glucose Traces of a Clinical Trial," *Diabetes Technology and Therapeutics*, vol. 16, no. 7, pp. 428–434, Jul 2014.

[169] Office for National Statistics, "Coronavirus (COVID-19) Latest Insights," *UK Office for National Statistics*, Jan 2023.

[170] L. Charleux, E. Roux, T. Goyallon, G. Feverati, and P. Nagorny, "Python Examples for Lotka-Volterra Equations," *Scientific Python*, 2021.

[171] J. Xie, "Simglucose," *Python Package*, 2018.

[172] B. Settles and M. Craven, "An Analysis of Active Learning Strategies for Sequence Labeling Tasks," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2008, p. 1070–1079.

[173] D.-A. Clevert, T. Unterthiner, and S. Hochreiter, "Fast and Accurate Deep Network Learning by Exponential Linear Units (ELUs)," *International Conference on Learning Representations (ICLR)*, 2016.

[174] A. F. Agarap, "Deep Learning using Rectified Linear Units (ReLU)," *arXiv*, vol. 1803.08375, 2018.

[175] A. L. Maas, "Rectifier Nonlinearities Improve Neural Network Acoustic Models," in *Proceedings of the 30th International Conference on Machine Learning*, 2013.

[176] G. Klambauer, T. Unterthiner, A. Mayr, and S. Hochreiter, "Self-Normalizing Neural Networks," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, 2017.

[177] D. Hendrycks and K. Gimpel, "Gaussian Error Linear Units (GELUs)," *arXiv*, vol. 1606.08415, 2023.

[178] S. Narayan, "The Generalized Sigmoid Activation Function: Competitive Supervised Learning," *Information Sciences*, vol. 99, no. 1, 1997.

[179] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," *International Conference on Learning Representations*, 2015.

[180] M. D. Zeiler, "ADADELTA: An Adaptive Learning Rate Method," *arXiv*, vol. 1212.5701, 2012.

[181] J. Duchi, E. Hazan, and Y. Singer, "Adaptive Subgradient Methods for Online Learning and Stochastic Optimization," *Journal of Machine Learning Research (JMLR)*, vol. 12, 2011.

[182] G. Hinton, N. Srivastava, and K. Swersky, "RMSProp," *Neural Networks and Machine Learning*, 2014.

[183] P. K. Mallapragada, R. Jin, A. K. Jain, and Y. Liu, "SemiBoost: Boosting for Semi-Supervised Learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 11, pp. 2000–2014, 2009.

[184] J. Mukhoti, V. Kulharia, A. Sanyal, S. Golodetz, P. H. S. Torr, and P. K. Dokania, "Calibrating Deep Neural Networks using Focal Loss," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, pp. 15 288–15 299, 2020.

[185] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal Loss for Dense Object Detection," *IEEE International Conference on Computer Vision (ICCV)*, pp. 2999–3007, 2017.

[186] A. Mavrin, "Focal Loss TensorFlow Implementation," *Python Package*, 2019.

[187] Q. Wu and E. Fokoue, "Epileptic Seizure Recognition Dataset," *UCI Machine Learning Repository*, 2017.

[188] M. Kachuee, S. Fazeli, and M. Sarrafzadeh, "ECG Heartbeat Classification: A Deep Transferable Representation," *IEEE International Conference on Healthcare Informatics (ICHI)*, pp. 443–444, 2018.

[189] S. Fazeli, "ECG Heartbeat Categorization Dataset," *Kaggle*, 2018.

[190] G. B. Moody and R. G. Mark, "The Impact of the MIT-BIH Arrhythmia Database," *IEEE Engineering in Medicine and Biology*, vol. 20, no. 3, pp. 45–50, 2001.

[191] M. Malekzadeh, R. G. Clegg, A. Cavallaro, and H. Haddadi, "Mobile Sensor Data Anonymization," in *Proceedings of the International Conference on Internet of Things Design and Implementation (IoTDI)*, 2019, p. 49–58.

[192] M. Malekzadeh, "Motion Sense Dataset," *GitHub*, 2019.

[193] T. Danka and P. Horváth, "modAL: A Modular Active Learning Framework for Python," *arXiv*, vol. 1805.00979, 2018.

[194] M. Feurer, A. Klein, K. Eggensperger, J. T. Springenberg, M. Blum, and F. Hutter, "Auto-sklearn: Efficient and Robust Automated Machine Learning," *Springer Automated Machine Learning*, pp. 113–134, 2019.

[195] T. O'Malley, E. Bursztein, J. Long, F. Chollet, H. Jin, and L. Invernizzi, "Keras Tuner," *Keras*, 2019.

[196] T. Xia, J. Han, and C. Mascolo, "Benchmarking Uncertainty Quantification on Biosignal Classification Tasks Under Dataset Shift," *Springer Multimodal AI in Healthcare*, p. 347–359, Nov 2022.

[197] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, 2017.

[198] H. Yao, C. Choi, B. Cao, Y. Lee, P. W. W. Koh, and C. Finn, "Wild-Time: A Benchmark of In-the-Wild Distribution Shift over Time," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 35, pp. 10 309–10 324, 2022.

[199] Y. Yang, H. Zhang, D. Katabi, and M. Ghassemi, "Change is Hard: A Closer Look at Subpopulation Shift," in *Proceedings of the 40th International Conference on Machine Learning (ICML)*, vol. 202. PMLR, Jul 2023, pp. 39 584–39 622.

[200] A. Simons, T. Doyle, D. Musson, and J. Reilly, "Impact of Physiological Sensor Variance on Machine Learning Algorithms," *IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pp. 241–247, 2020.

[201] D. Spathis, I. Perez-Pozuelo, T. I. Gonzales, Y. Wu, S. Brage, N. Wareham, and C. Mascolo, "Longitudinal Cardio-Respiratory Fitness Prediction through Wearables in Free-Living Environments," *npj Digital Medicine*, vol. 5, no. 1, Dec 2022.

[202] M. Chen, L. Shen, H. Fu, Z. Li, J. Sun, and C. Liu, "Calibration of Time-Series Forecasting: Detecting and Adapting Context-Driven Distribution Shift," in *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*. Association for Computing Machinery, 2024, p. 341–352.

[203] R. De Maesschalck, D. Jouan-Rimbaud, and D. Massart, "The Mahalanobis Distance," *Chemometrics and Intelligent Laboratory Systems*, vol. 50, no. 1, pp. 1–18, 2000.

[204] A. Venkataramanan, A. Benbihi, M. Laviale, and C. Pradalier, "Gaussian Latent Representations for Uncertainty Estimation using Mahalanobis Distance in Deep Classifiers," *IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pp. 4490–4499, Oct 2023.

[205] S. Govindaraj and T. Tejas, "The Hellinger Distance and its Applications to Hypothesis Testing and Model Uncertainty," *SSRN Electronic Journal*, 2022.

[206] Y. Zheng, F. Yang, J. Duan, and J. Kurths, "Quantifying Model Uncertainty for the Observed Non-Gaussian Data by the Hellinger Distance," *Communications in Nonlinear Science and Numerical Simulation*, vol. 96, p. 105720, 2021.

[207] P.-E. Danielsson, "Euclidean Distance Mapping," *Computer Graphics and Image Processing*, vol. 14, no. 3, pp. 227–248, 1980.

[208] Y. Gal and Z. Ghahramani, "Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning," in *Proceedings of The 33rd International Conference on Machine Learning*, vol. 48. PMLR, Jun 2016, pp. 1050–1059.

[209] E. J. Hu, yelong shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "LoRA: Low-Rank Adaptation of Large Language Models," *International Conference on Learning Representations (ICLR)*, 2022.

[210] S. Kiranyaz, O. Avci, O. Abdeljaber, T. Ince, M. Gabbouj, and D. J. Inman, "1D Convolutional Neural Networks and Applications: A Survey," *Mechanical Systems and Signal Processing*, vol. 151, p. 107398, 2021.

[211] G. Clifford, C. Liu, B. Moody, L.-w. Lehman, I. Silva, Q. Li, A. Johnson, and R. Mark, "AF Classification from a Short Single Lead ECG Recording: the Physionet Computing in Cardiology Challenge 2017," *Computing in Cardiology Conference (CinC)*, Sep 2017.

[212] A. L. Goldberger, L. A. N. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley, "PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals," *Circulation*, vol. 101, no. 23, Jun 2000.

[213] G. Clifford, C. Liu, B. Moody, L.-w. Lehman, I. Silva, Q. Li, A. Johnson, and R. Mark, "Physionet 2017 ECG," *Kaggle*, 2017.

[214] G. Moody and R. Mark, "MIT-BIH Arrhythmia Database," *PhysioNet*, 2005.

[215] R. Bousseljot, D. Kreiseler, and A. Schnabel, "Nutzung der EKG-Signaldatenbank Cardiodat der PTB über das Internet," *De Gruyter Brill*, vol. 40, pp. 317–318, 1995.

[216] R. Bousseljot, "PTB Diagnostic ECG Database," *PhysioNet*, 2004.

[217] B. Moody, G. Moody, M. Villarroel, G. Clifford, and I. Silva, "MIMIC-III Waveform Database," *PhysioNet*, Apr 2020.

[218] P. H. Charlton, K. Kotzen, E. Mejía-Mejía, P. J. Aston, K. Budidha, J. Mant, C. Pettit, J. A. Behar, and P. A. Kyriacou, "Detecting Beats in the Photoplethysmogram: Benchmarking Open-Source Algorithms," *Physiological Measurement*, vol. 43, no. 8, p. 085007, Aug 2022.

[219] P. H. Charlton, "MIMICPERform Datasets," *Zenodo*, Aug 2022.

[220] A. H. Ribeiro, M. H. Ribeiro, G. M. M. Paixão, D. M. Oliveira, P. R. Gomes, J. A. Canazart, M. P. S. Ferreira, C. R. Andersson, P. W. Macfarlane, W. Meira Jr., T. B. Schön, and A. L. P. Ribeiro, "Automatic Diagnosis of the 12-Lead ECG Using a Deep Neural Network," *Nature Communications*, vol. 11, no. 1, p. 1760, 2020.

[221] A. H. Ribeiro, M. H. Ribeiro, G. M. Paixão, D. M. Oliveira, P. R. Gomes, J. A. Canazart, M. P. Ferreira, C. R. Andersson, P. W. Macfarlane, W. Meira Jr., T. B. Schön, and A. L. P. Ribeiro, "CODE-test: An Annotated 12-Lead ECG Dataset," *Zenodo*, Jan 2020.

[222] A. de Mathelin, M. Atiq, G. Richard, A. de la Concha, M. Yachouti, F. Deheeger, M. Mougeot, and N. Vayatis, "ADAPT: Awesome Domain Adaptation Python Toolbox," *arXiv*, vol. 2107.03049, 2023.

[223] H. Daumé III, "Frustratingly Easy Domain Adaptation," in *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, 2007, pp. 256–263.

[224] M. Loog, "Nearest Neighbor-Based Importance Weighting," in *2012 IEEE International Workshop on Machine Learning for Signal Processing*, 2012, pp. 1–6.