UNIVERSITY OF CAMBRIDGE

Toward effective and generalisable machine learning for biosignal time series

Yu Wu

Robinson College

September 2025

This thesis is submitted for the degree of *Doctor of Philosophy* at the Department of Computer Science and Technology

Declaration

This thesis is the result of my own work and includes nothing which is the outcome of work done in collaboration 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

Biosignal time series collected from wearable devices, such as electrocardiograms (ECG), electroencephalograms (EEG), and Inertial Measurement Units (IMUs), enable continuous monitoring of human physiology and behaviour. Using these signals provides unique opportunities to advance personalised health monitoring, improve early disease detection, and support the development of scaleable and cost-effective healthcare solutions. However, analysing such data presents significant challenges due to its complex nature, as signals are often sparse, irregular, contain missing values, and lack sufficient labels. In contrast, existing state-of-the-art methods are typically developed and validated on clean, regularly sampled biosignals with clinically verified ground-truth annotations, which limits their effectiveness and leads to performance degradation when applied to real-world data. Moreover, most existing models for biosignal time series are developed and evaluated in-domain, where both training and testing are conducted on the same dataset or under the same data collection protocol. Therefore, such models often fail to generalise to new scenarios in real-world healthcare deployments due to distributional shifts across tasks or deployment settings.

This thesis addresses these challenges by developing machine learning methods tailored for biosignal time series that focus on self-supervised representation learning, domain adaptation, and foundation modelling for generalisation across tasks and datasets. We demonstrate the efficacy of these methods on a wide range of real-world healthcare tasks, from daily cardio-fitness monitoring to clinical applications.

First, to address the scarcity of labelled data and the inherent complexity of biosignal time series, we propose a contrastive self-supervised learning framework specifically designed for biosignals, collected across diverse sources, ranging from wearable devices to clinical monitoring systems. Inspired by the success of contrastive learning in computer vision, our method captures the temporal and structural characteristics of biosignals without relying on labels. Then, our new training pipeline learns more effective representations from unlabelled data and improves the downstream task performance. This performance remains comparable to state-of-the-art methods, even when only a small fraction of labelled data is available, thus enhancing label efficiency.

Second, to improve model robustness under distribution shifts, we develop a domain

adaptation framework with multiple discriminators during fine-tuning. This method learns domain-invariant representations that generalise across source and target domains with differing label distributions, which is a common issue in healthcare due to the high cost and scarcity of gold-standard annotations. We validate this approach on a VO_2max prediction task and demonstrate improved adaptation performance under real-world domain shifts.

Finally, motivated by the growing success of foundation models in language and vision, we introduce a general-purpose foundation model tailored for biosignal time series to alleviate data complexity and improve model generalisability. Our approach pretrains on heterogeneous datasets that vary in modality, sampling rates, and missingness patterns, and is capable of generalising to a wide range of unseen downstream tasks, from classification to regression, across different biosignal sources. This foundation model provides a scaleable and flexible framework for developing future biosignal-based healthcare applications.

Together, this dissertation tackles three central challenges in biosignal time series modelling: (i) data and label scarcity, addressed through contrastive self-supervised learning for label-efficient representation learning; (ii) domain shifts, mitigated by a multi-discriminator domain adaptation framework; and (iii) data and domain heterogeneity, alleviated by a general-purpose biosignal foundation model. These contributions were validated on real-world physiological datasets as well as commonly used machine learning benchmarks, with extensive experiments demonstrating their effectiveness in handling missingness, irregularity, and limited labels while enabling generalisable model deployment across diverse healthcare scenarios. Collectively, this progression from data-efficient modelling to robust domain adaptation and ultimately to scaleable generalisation underscores the potential of deep learning to advance the practical adoption of biosignal-based healthcare solutions.

Acknowledgements

My PhD journey began in the autumn of 2021, and now, four years later in the autumn of 2025, I finally find myself writing these acknowledgements. I once imagined that this moment would be overwhelming, that I would be moved to tears. Yet, now that it has arrived, I struggle to find the right words, the right tone, or even the right place to begin. Over these years I have gradually learned how to write academic papers. But unlike academic writing, this page carries far too many people to thank, and far too many moments worth remembering. I only hope that my words do justice to them all.

First and foremost, I owe my deepest gratitude to my advisor, Professor Cecilia Mascolo, for her exceptional guidance, unwavering support, and invaluable mentorship throughout my PhD. Her quick thinking, rigorous logic, and sharp insights into problems have shaped the researcher I am today. Her working efficiency, tireless support for her students, and dedication to building and leading the Mobile Systems Group are qualities I will always admire and learn from. I am profoundly grateful for the environment she created, where I had the opportunity to grow alongside her and my colleagues.

I am also deeply grateful to my talented collaborators, Dr. Ting Dang, Dr. Dimitris Spathis, and Dr. Hong Jia, who played a crucial role in making this thesis possible. I cannot thank them enough for their passion, dedication, and generosity, which constantly inspired me in research. I will always remember Ting's meticulous guidance and her willingness to spend endless hours in discussion, as well as her perseverance as a researcher. Dimitris's enthusiasm and his ability to view research problems from broad and diverse perspectives have continually motivated me. Hong's sharp instincts and detailed feedback on experimental methods were invaluable to my training. Without them, I could not have stayed on the right path in my research journey.

My internship at Microsoft Research Asia was a truly remarkable experience that opened a new chapter in my research journey and broadened my horizons. Working with Professor Lili Qiu's team offered me invaluable opportunities to exchange ideas, collaborate across diverse research backgrounds, and persist through trials and failures together. The collaborative spirit and stimulating environment helped me consolidate my research interests and develop a clearer vision for my future work.

Furthermore, I am grateful to my examination committee, Professors Mateja Jamnik

and Aaqib Saeed, for their patience in reading this thesis—likely the only two people in the world who have gone through it line by line. I deeply appreciate the insightful discussions during the viva and their thoughtful suggestions, which have greatly improved the quality of my work.

Being part of the Mobile Systems Group has made these years both enjoyable and memorable. I would like to thank Dr. Tong Xia for our many academic discussions, Dr. Yang Liu and Dr. Qiang Yang for generously sharing their experiences and inspiring me with their dedication to research. I am equally thankful to Evelyn, Jing, Kayla, Sotiris, Jake, Mathias, and Young—I will never forget our daily fun, discussions, and countless meals together. Research is never done in isolation, and I am truly grateful to my friends in Cambridge, UK and abroad who have brightened my life, listened to my random thoughts, and shared both meaningful and silly conversations with me. Their companionship has been a strong source of support.

Finally, to my parents, for their unconditional love, constant understanding, and endless support. I am forever inspired by their kindness, warmth, and devotion to their work. They have always stood behind me, and my gratitude to them can never be fully expressed. I love them deeply.

People often say that pursuing a PhD is a lonely journey that requires enormous effort. I agree. But along the way, I have been fortunate to receive the support from brilliant mentors, dear friends, and my family. My final tribute is to all of them, who know that I am not perfect but still chose to show up for me, listen to me, care for me, and love me. Beyond the research career, it is their character, the landscapes of this journey, and the moments we shared that gave me strength, courage, curiosity, and the desire to continue exploring and wandering without end.

Contents

Chapter 1

Introduction

1.1 Motivation

The growing demand for healthcare services, driven by the global demographic shifts of a population ageing and the increasing prevalence of chronic diseases, is placing unprecedented pressure on health systems worldwide. By 2030, the number of people aged 60 and older is projected to reach 1.4 billion, up from 1 billion in 2020, substantially intensifying the strain on clinical resources (World Health Organization, 2022). Despite these growing needs, traditional healthcare systems, which are largely reliant on institutional episodic care, provide timely, equitable, and sustainable care that meets the needs of increasingly diverse populations.

In response, smart health systems have emerged as a promising paradigm for extending healthcare beyond clinical settings. Leveraging advances in sensing technologies and artificial intelligence (AI), these systems enable continuous and real-time monitoring of physiological states, supporting preventive care, early diagnosis, and personalised intervention (Baig and Gholamhosseini, 2013; Ghulam et al., 2021). Central to these systems are biosignal time series, which are continuous physiological signals such as electrocardiography (ECG), electroencephalography (EEG), photoplethysmography (PPG) and inertial measurement units (IMUs). With the widespread adoption of consumer-grade wearable devices, it has become possible to routinely collect biosignals in daily life, allowing real-time monitoring of physical activity, stress, sleep, and early signs of disease (Jahmunah et al., 2019; Dauwels et al., 2010; Kwon et al., 2021).

In parallel, recent years have seen a boom in machine learning (ML) techniques, which have demonstrated impressive results in healthcare using the power of data analysis and modelling (Ahmad et al., 2018). For example, a convolutional neural network trained on skin lesion images was able to classify skin cancers with an accuracy that is on par with board certified dermatologists (Esteva et al., 2017). Such dermatologist-level image recognition illustrates how AI can assist or even enhance traditional diagnostics.

Likewise, large language models (LLMs) such as Google's Med-PaLM (Singhal et al., 2022) have demonstrated the ability to become the first form of AI to pass the U.S. Medical Licensing Exam and can provide high-quality answers to medical questions. These advances, spanning from computer vision to an understanding of natural language understanding, have highlighted the immense potential of AI to improve disease detection, diagnosis, and healthcare delivery. They offer scaleable tools to improve clinician workflows and improve health outcomes.

However, translating this progress to biosignal time series analysis remains an open and as yet underdeveloped frontier. In fact, there remains a huge gap between the existing time series literature and what is needed to make ML systems practical and deployable for healthcare (Loni et al., 2024). Unlike structured tabular data or static medical images, biosignal time series present a unique set of challenges in that they are often sparse, irregularly sampled, multi-modal, and contain missing values. Labels are expensive to obtain, frequently noisy, or entirely absent. Furthermore, the distribution of data can shift significantly over time or across populations, thereby limiting the generalisability of the model. These factors make model training and deployment particularly difficult and explain why, despite growing academic interest and the proliferation of wearables, only a handful of ML-based biosignal applications have reached the clinical adoption threshold (Sabry et al., 2022). For example, the irregular heart rhythm notification algorithm on the Apple Watch is one of the few successes, having received FDA approval in 2018 to alert users to potential atrial fibrillation (Sabry et al., 2022). However, the replicability and reproducibility of the algorithm can be hampered moving forward due to the lack of standardisation, and this can affect the generalisability of the findings (Nelson et al., 2020). In general, the practical, effective, and reliable use of ML for biosignal time series still faces many obstacles. A detailed discussion of how these factors present substantial obstacles to our goal is elaborated in Section 1.2.

This thesis aims to bridge this gap by developing ML models that are effective, generalisable, and tailored to the real-world challenges of biosignal time series. Through novel approaches in representation learning, domain adaptation, and foundational modelling, it contributes both algorithmic insights and application-driven solutions. Ultimately, this work aims to advance the deployment of cost-effective, reliable, and intelligent health monitoring systems to bring us closer to a future of proactive and personalised healthcare.

1.2 Limitation and research questions

Developing effective and generalisable machine learning models for biosignal time-series data in healthcare poses multiple challenges. These challenges span the entire modelling pipeline. They include limitations in raw data quality and annotation, difficulties in

deploying across diverse domains, and the lack of unified foundation models for complex healthcare signals. This thesis specifically addresses these challenges with the aim of building models that are robust, adaptable, and scaleable for health monitoring applications.

i.) Real-World biosignal time series are complex and difficult to annotate.
The increased availability of biosignal time series has led to a growing interest in data-driven modelling. Data collected from common biosignal sensors (*e.g.*, accelerometers for motion behavioural monitoring, or ECGs for cardiac activity) are commonly represented as biosignal time series. Unlike static data modalities such as images or text, biosignal time series are fundamentally temporal in nature, requiring models to capture sequential patterns and dynamic dependencies over time. The ability to effectively model such temporal data is crucial not only for identifying patterns inherent in physiological processes but also for enabling accurate forecasting, anomaly detection, and prediction of health outcomes.

However, such data are inherently noisy, often irregularly sampled, and frequently suffer from missing values due to sensor dropouts or user non-compliance (Weerakody et al., 2021). These issues complicate the development of robust ML models. Compounding the data quality challenge is the scarcity of high-quality labels. Annotating health outcomes (*e.g.*, VO_2max, stress levels, or mental states) typically requires expensive and labour intensive procedures involving laboratory-grade equipment, expert annotation, or clinical evaluations (Uth et al., 2004). This makes it impractical to scale supervised learning approaches, especially when continuous monitoring is required in free-living conditions.

Traditional models trained on small, fully labelled datasets often underperform or overfit, thus limiting their utility in real-world applications (Ying, 2019). These challenges motivate the need for learning paradigms that can extract meaningful representation from minimally labelled data, which are capable of leveraging structure from the data itself, rather than relying solely on human-provided annotations.

ii.) Health models struggle to generalise to a new inference domain.
In healthcare settings, ML models must be reliable and generalisable, particularly because they are expected to perform well, not only on training data, but also in unseen domains [1] (Zadorozhny et al., 2022). This is especially important in practice, where models are often trained under controlled laboratory conditions but are expected to operate in a variety of real-world contexts, devices, and patient populations. Despite extensive research and the development of numerous models, relatively few time series health applications

[1] We define *domain* to refer to a specific dataset or data distribution defined by factors such as sensor modality, acquisition environment, population demographics, or labelling protocol. *Cross-domain* refers to scenarios in which training and evaluation data are drawn from different distributions. Such differences may arise from variations in labelling protocols (leading to label distribution shifts), discrepancies in the number or type of sensors, or differences in missing data patterns and sampling characteristics.

driven by ML have reached widespread deployment (Qayyum et al., 2020).

One of the key challenges limiting broader adoption is the issue of distribution shift, which occurs when changes in sensor heterogeneity, user behaviour, clinical conditions, or label granularity cause differences in data distributions between training and deployment. For example, models trained on ECG data from a clinical setting may not generalise to signals collected via wearables during daily activity, due to differences in positioning, or signal morphology. Many existing studies report high performance when training and testing are conducted on curated datasets with identical distributions. However, these same models often fail to generalise when deployed in new environments. Without mechanisms to adapt to such distribution shifts, models exhibit degraded performance in new domains, ultimately limiting their practical utility and reliability in healthcare settings. Addressing this requires domain-adaptive techniques and robust modelling strategies that can learn domain-invariant features and maintain performance across heterogeneous devices, sensing protocols, and deployment contexts.

iii). Lack of a foundation model to deal with complex real-world data and cross-domain generalisation.

Foundation models represent a new paradigm in ML in which large-scale models trained are on broad, diverse datasets using self-supervised learning to learn general-purpose representations (Liang et al., 2024). Unlike traditional supervised models that rely on large annotated datasets, foundation models learn general-purpose representations by leveraging unlabelled data, commonly through techniques such as masked prediction, where certain parts of the data are masked and the model is asked to predict them. This approach enables the model to distill complex patterns from the data and form a pretrained core that can be adapted to a wide range of downstream tasks. Foundation models have demonstrated remarkable success in fields such as natural language processing (NLP; *e.g.*, GPT and BERT) (Achiam et al., 2023; Koroteev, 2021) and computer vision (*e.g.*, CLIP and DINO) (Hafner et al., 2021; Zhang et al., 2022a), where they have enabled rapid adaptation to new tasks with minimal supervision. For example, in speech recognition, self-supervised foundation models (*e.g.*, Wav2Vec) have dramatically reduced the need for labelled audio by pretraining on large unlabelled datasets and then fine-tuning for specific speech tasks (Baevski et al., 2020).

In contrast, biosignal time series have yet to fully benefit from such advancements. Current deep learning models for biosignals are typically handcrafted for specific tasks or datasets and trained on relatively small, curated collections of data (Wan et al., 2023). However, these models often assume idealised conditions. For example, signals are clean, regularly sampled, and confined to a single modality, which rarely holds in real-world settings. In practice, biosignal data from wearables or medical monitors are inherently

noisy, often irregularly sampled, and prone to missing values due to sensor dropouts or user non-compliance. Variations between data sources (different subjects, devices, or environments) also introduce domain shifts that challenge model generalisation.

Because existing models are narrowly tailored and lack robust pretraining on diverse biosignals, they tend to struggle with missingness, irregularity, and cross-domain generalisation. In other words, without a foundation model in this domain, each new sensing modality or health application often requires developing a model from scratch and extensive re-engineering for that context. This gap motivates the need for biosignal foundation models, large-scale pretrained models that can handle real-world complexity and variability and which can be adapted to downstream health tasks with minimal labelled data.

1.3 Thesis and substantiation

We have reviewed the promising opportunities enabled by ML advancements in biosignal modelling, as well as the fundamental challenges involved in developing effective models for complex, real-world data. Formally, the overarching objective of this thesis can be stated as: *To design effective ML models for biosignal time-series data with the goal of generalising to unseen datasets and tasks in order to improve health monitoring and human well-being.* We substantiate this statement by first evaluating the potential of existing approaches on biosignal time series datasets and then proposing new models that outperform current methods or offer new insights. Our methods leverage and expand on the paradigms of domain adaptation and self-supervised learning, culminating in a foundation model. In particular, this dissertation addresses the following research questions:

- **Research Question 1.** How can we use ML to learn representations from unlabelled biosignal time-series data and perform effectively and achieve strong performance on diverse classification tasks?

- **Research Question 2.** How can we generalise models trained on large-scale, free-living biosignal data with noisy or weak labels to accurately predict gold-standard health outcomes?

- **Research Question 3.** How can we build general-purpose foundation models that handle the complexities of biosignal time-series data—including irregular sampling, missingness, and modality heterogeneity—while generalising effectively across domains and downstream healthcare tasks?

To address these questions, we develop novel self-supervised models which can tackle complex biosignal time series, such as labelling issues, and missingness and irregularity. Further, we design novel transfer learning that leverages a large pretrained model and is

able to generalise to new domain to predict different healthcare tasks. Finally, building on these advances, we introduce a general-purpose biosignal foundation model that unifies data-efficient representation learning and cross-domain adaptability, providing a scaleable framework for diverse real-world healthcare applications.

1.4 Contribution and chapter outline

In terms of methods, this thesis customises extant ML models, particularly deep learning architectures, to the unique characteristics of biosignal time series data. On the application side, it addresses a spectrum of challenges in both clinical and free-living healthcare settings. Chapter 2 provides an overview of the background and related work in deep learning for biosignal time series, followed by three core contributions presented in the subsequent chapters:

Contribution 1: Self-supervised learning for unlabelled biosignal time series data

Chapter 3 investigates how self-supervised learning, specifically contrastive learning (CL) can be an effective way to extract robust representation from unlabelled biosignal time series. CL adopts the idea of attracting similar samples and distracting dissimilar samples to solve the unlabelled issue. However, the existing CL for biosignal time series often neglects the inherent characteristics of time series. As a result, they suffer from false negative pairs (FNPs), where semantically or temporally similar signals are mistakenly treated as dissimilar, thus degrading performance and label efficiency. To address this, we propose StatioCL, a novel contrastive learning framework that systematically mitigates these FNPs.

StatioCL identifies and categorises two types of false negatives that are particularly problematic in biosignal learning, specifically semantic FNPs, which arise when similar physiological states (*e.g.*, resting heart rate or similar movement patterns) are incorrectly treated as dissimilar; and temporal FNPs, which occur when temporally adjacent signal segments from the same context are mistakenly contrasted. These are prevalent in biosignal data due to non-stationarity and temporal correlations in health and activity patterns. To mitigate these issues, StatioCL introduces two key contributions: A *non-stationary state encoder* that identifies and encodes segment-level trends and dynamic changes over time, allowing the model to avoid contrasting semantically similar states. A *temporal contrast mechanism* that computes fine-grained similarity scores based on segment proximity and dependencies, reducing temporal false negatives during contrastive learning.

We validate StatioCL on multiple real-world biosignal benchmarks and show that it consistently outperforms existing state-of-the-art contrastive learning baselines. StatioCL

improves recall by 2.9% on average and reduces the incidence of false negative pairs by 19.2%. Moreover, the model exhibits enhanced data efficiency and robustness to label scarcity, performing competitively even when trained with fewer labelled examples. This work contributes to the development of more accurate and efficient self-supervised models for biosignal time series, addressing a core limitation in current contrastive learning pipelines and paving the way for generalisable representation learning in health monitoring contexts.

Contribution 2: Domain adaptation to alleviate noisy labels to improve cardiorespiratory fitness prediction via free-living wearable devices

Chapter 4 explores how domain adaptation techniques enable ML models to generalise under label distribution shift, an increasingly common and challenging scenario in real-world healthcare applications. In practice, high-quality (or gold-standard), clinically validated datasets are typically small-scale and thereby difficult to scale, whereas imprecise (or silver-standard) datasets derived from wearables are increasingly available at population scale but suffer from label noise and distribution mismatches.

To address this, we propose UDAMA, an unsupervised domain adaptation framework that leverages large-scale wearable data with silver-standard labels to improve model performance on gold-standard targets. During the pretraining phase, UDAMA learns an initial representation by training a feature extractor on large-scale silver-standard data with noisy labels. This phase allows the model to learn population-level signal patterns from data collected under free-living conditions, despite imprecision of the labels. In the adaptation (fine-tuning) phase, UDAMA introduces a multi-discriminator adversarial learning strategy to align the distributions of feature representations between the source (silver-standard) and target (gold-standard) domains. Specifically, it deploys two discriminators: one operating on the raw feature embeddings and the other on the predicted label space. The dual discriminator design ensures that both the intermediate representation and the output predictions are domain-invariant, thereby mitigating distribution shifts between datasets.

This chapter focusses on the practical task of predicting cardiorespiratory fitness (CRF), an important clinical marker of metabolic health and mortality risk. CRF is difficult to estimate directly in free-living settings due to the cost of ground-truth measurements such as VO_2max. Alternatively, it can be indirectly and imprecisely assessed using the heart rate response to a standard exercise test to evaluate the VO_2max. Subsequently, we evaluated UDAMA using two wearable cohorts on a population scale to pre-train and test our model. The first of these is the Fenland Study, with 11,059 participants and algorithm-derived silver-standard VO_2max labels, and the second is the BBVS cohort, which includes 281 participants with high-quality labels from exercise tests. We then used 60 participants from the BBVS with gold-standard labels as the hold-out test set.

Our model achieves a person-level correlation of 0.701 ± 0.032 on the gold standard test set, surpassing even supervised baselines trained and evaluated within the same domain. Furthermore, UDAMA consistently outperforms conventional transfer learning and state-of-the-art domain adaptation baselines by up to 12% and demonstrates robustness to a variety of synthetic label shift scenarios. This work contributes to the development of scaleable and generalisable digital health systems by showing how domain adaptation can effectively bridge noisy and high-fidelity data sources, thus enabling robust, clinically relevant fitness estimations to be derived from large-scale wearable signals.

Contribution 3: Foundation model for multi-domain irregular time series
Chapter 5 presents a foundation model approach to learning from irregularly sampled, missing, and multi-domain time series data, a common characteristic in wearable biosignal monitoring across clinical and real-world settings. The proposed framework aims to generalise across multiple domains, sensors, and missingness patterns, thereby enabling the scaleable regression modelling of physiological and behavioural signals with minimal supervision.

This chapter begins by reviewing existing foundation models, which are often trained in a self-supervised manner for time series and are usually focused on a single modality, and thus lack the ability to generalise across domains or datasets. More importantly, existing foundation models still assume that training time series data are clean, meaning continuously and regularly sampled without missingness, even if such data is multivariate and aligned, which is rare in the healthcare settings and thus reveals performance degradation on regular biosignal time series datasets.

To address this, we introduce FORM, a foundation model trained from multi-domain, irregular time series. FORM is designed to generalise across datasets with varying numbers of sensors, sampling patterns, and missing rates. At the core of FORM is a self-supervised pretraining framework based on masked reconstruction, designed for real-world biosignals that are often noisy, irregular and multimodal. Specifically, FORM explicitly learns to reconstruct masked tokens from irregularly sampled sequences using an irregularity-sensitive masking strategy. The masking scheme is designed to preserve the temporal structure and irregularity of real-world signals, thereby encouraging the model to learn patterns that are robust to missingness and irregular sampling. To further support cross-domain generalisation and flexible sensor configurations, FORM adopts a channel-independent encoding strategy. Each sensor channel is encoded independently using a shared encoder, avoiding assumptions about channel alignment or co-sampling. This allows the model to scale across datasets with heterogeneous sensing modalities and varying sensor counts, without need for architectural changes or retraining.

We evaluated the framework in several irregular biosignal datasets that span different

domains, from intensive care unit (ICU) settings to daily human activity recognition, sensor modalities, and temporal irregularity characteristics. The results show that FORM achieves strong generalisation between domains, especially in low data regimes and under different missingness conditions. The model outperforms the previous baselines by up to 6.48% in irregular time series regression and by 3.5% in irregular time series classification tasks, even with minimal fine-tuning.

This work contributes a step toward building foundation models for time series, offering a unified and flexible framework for modelling biosignals in the presence of noise, irregular sampling, and cross-domain variability. Consequently, it lays the foundation for scaleable and generalisable digital health solutions across a wide range of sensing platforms and health indicators.

1.5 List of publications

During my PhD studies, I was fortunate to establish several collaborations with computer scientists, engineers, and other domain experts, which have yielded publications both in ML methods and applications to healthcare. In particular, Chapter 3 draws from a study published in CIKM 2024 (Wu et al., 2024), Chapter 4 is based on a paper at NeurIPS ML4H 2022 (Wu et al., 2022) and MLHC 2023 (Wu et al., 2023), and finally Chapter 5 is based on a work under review. Beyond this, I co-authored some other works in the wider area of machine learning and mobile sensing, while not directly related to this thesis, which have nevertheless influenced my ideas.

Work related to this dissertation

- **Wu, Y.**, Dang, T., Spathis, D., Jia, H., & Mascolo, C. (2024). StatioCL: Contrastive Learning for Time Series via Non-Stationary and Temporal Contrast. *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*. https://doi.org/10.1145/3627673.3679732

- **Wu, Y.**, Spathis, D., Jia, H., Perez-Pozuelo, I., Gonzales, T.I., Brage, S., Wareham, N. &; Mascolo, C. (2023). UDAMA: Unsupervised Domain Adaptation through Multi-discriminator Adversarial Training with Noisy Labels Improves Cardio-fitness Prediction. *Proceedings of the 8th Machine Learning for Healthcare Conference, in Proceedings of Machine Learning Research* 219:863-883, https://proceedings.mlr.press/v219/wu23a.html.

- **Wu, Y.**, Spathis, D., Jia, H., Perez-Pozuelo, I., Gonzales, T.I., Brage, S., Wareham, N.J., & Mascolo, C. (2022). Turning Silver into Gold: Domain Adaptation with

Noisy Labels for Wearable Cardio-Respiratory Fitness Prediction. *Machine learning for Health* ArXiv, abs/2211.10475.

- **Wu, Y.**, Zheng, F., Dang, T. & Mascolo, C. (2025). FORM: A Foundation Model for Irregular Multi-domain Time Series (under review).

Other works and preprints

- **Wu, Y.***, Chen, X.*, Lee, S.*, Yoon, H.*, Lu, T., Liu, Y., Lee, S., Chen, D., Mascolo, C, & Qiu, L. (2025). NeuroBuds: Towards a Foundation Model for Task-agnostic ExG Sensing from Earphones (under review). *Equal contribution

- Zheng, F., **Wu, Y.**, Mascolo, C, & Dang, T. (2025). Rethinking Large Language Models for Time Series Classification in Critical Care (under review).

- Vavaroutas, S., **Wu, Y.**, Etemad, A., & Mascolo, C. (2025). ADAPTOOD: Uncertainty-Aware Fine-Tuning for Out-of-Distribution ECG Time Series Models (under review).

- Dang, T., Chatterjee, S., Jia, H., **Wu, Y.**, Salim, F., & Kawsar, F. (2025): AdaN-ODEs: Test Time Adaptation for Time Series Forecasting Using Neural ODEs (under review).

- Zhang, Y., Xia, T., Han, J., **Wu, Y.**, Rizos, G., Liu, Y., Mosuily, M., Chauhan, J., & Mascolo, C. (2024). Towards Open Respiratory Acoustic Foundation Models: pretraining and Benchmarking. *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks track.*
https://doi.org/10.48550/arXiv.2406.16148

- Ketmalasiri, T., **Wu, Y.**, Butkow, K., Mascolo, C., & Liu, Y. (2024). IMChew: Chewing Analysis using Earphone Inertial Measurement Units. *Workshop on Body-Centric Computing Systems (BodySys @ MobiSys).*
https://dl.acm.org/doi/pdf/10.1145/3662009.3662023

- Spathis, D., Pozuelo, I., Gonzales, T., **Wu, Y.**, Brage, S., Wareham, N., & Mascolo, C. (2022). Longitudinal cardio-respiratory fitness prediction through wearables in free-living environments. *Nature Digital Medicine (npj Digit. Med.),5(176).*
https://doi.org/jpcc

Chapter 2

Background

Recent advances in machine learning, particularly in NLP and computer vision (CV), have significantly advanced the analysis of medical data and the development of diagnostic tools. Deep learning–powered health applications, such as disease pre-screening from biomedical data (Hemker et al., 2025) and self-diagnosis via medical chatbots (Habib et al., 2021), have demonstrated the transformative potential of these techniques. Motivated by these successes in diverse domains, this thesis focusses on deep learning-based approaches to modelling biosignal time series.

Building on the previous chapter, which highlights the importance of developing new models for biosignals using data collected from wearable devices and biosensors, this chapter outlines the necessary background and related literature. We begin with an introduction to biosignal time series with a view to describing their conceptualisation and characteristics (Section 2.1) and then provide an overview of the healthcare tasks and related datasets that are evaluated in Section 2.2. We then review the fundamental deep learning pipelines and core architectures (Section 2.3), before discussing the advanced training paradigms that state-of-the-art methods employ to address the unique data and labelling challenges inherent in biosignal modelling (Section 2.4).

2.1 Biosignal time series

Biosignals are measurements of physiological processes in living organisms, which may be recorded as electrical, mechanical, chemical, or optical signals (Schultz and Maedche, 2023). When measured continuously, they form biosignal time series that capture how physiological states change over time. Common examples include monitoring the electrical activity of the heart by ECG; recording of the activity of the brain via EEG; capturing inertial motion data from body movements by accelerometers or gyroscopes; or measuring changes in blood volume via photoplethysmography (PPG) (Bolpagni et al., 2024).

Specifically, vital signs form a clinically important subset of biosignals, representing

key physiological measurements that are essential indicators of life and immediate health status (Lee et al., 2022). These include heart rate, respiratory rate, blood pressure, and body temperature, with oxygen saturation often considered a fifth vital sign. Vital signs are typically derived from biosignals, such as the heart rate of ECG or PPG, and are routinely standardised for rapid assessment in routine check-ups and emergency care. Together, this wealth of such data and its derivatives can provide healthcare professionals with valuable information to make informed diagnostics and preventive decisions.

Historically, collecting such biosignals required specialised clinical equipment operated by trained professionals within controlled environments, limiting both the scale and diversity of the available data. However, rapid advances in mobile sensing and wearable technology have since changed this paradigm (Khan et al., 2016; Abbaspourazad et al., 2024). Today, it is possible to collect biosignals in free-living conditions, enabling continuous, large-scale, and real-world health monitoring. Examples of such technological developments include respiratory audio captured by microphones in earbuds (Liu et al., 2024), motion tracking via IMUs in smartwatches and smartphones (Auepanwiriyakul et al., 2020), and heart rate monitoring with PPG sensors embedded in wristbands or rings (Castaneda et al., 2018). This shift has dramatically expanded the potential applications of biosignal analysis, moving the field beyond the clinic and into everyday life.

While modern biosensors now allow biosignals to be collected under almost any condition, the resulting time series introduce unique analytical challenges. Unlike images or other static data types, biosignal time series often suffer from missing values, irregular sampling, and label scarcity due to sensor instability, battery limitations, or the difficulty of obtaining ground-truth health outcomes. For example, wearable biosignal data often contain gaps due to missing segments or exhibit noise caused by motion artefacts. Thus, obtaining medical annotations for long-term recordings is frequently impractical. Such data can also be highly heterogeneous and vary between devices, body locations, and individuals. Addressing these challenges is essential to unlock the full potential of biosignal data in real-world health monitoring and diagnostics.

These characteristics motivate the use of ML, which provides a powerful framework to discover patterns in complex, high-dimensional time series and predict health-related outcomes (Al-Turjman and Baali, 2022). In this thesis, we develop ML methods tailored to the unique characteristics of biosignal time series, with the aim of enabling robust and generalisable health monitoring. More details on model training are introduced in Section 2.3.

2.2 Healthcare tasks and biosignal datasets

In this thesis, we address a diverse range of healthcare tasks that leverage biosignal time series collected from both free-living and clinical environments. To provide a consistent view, we group these tasks into three broad categories, specifically, (i) physical activity and fitness monitoring, (ii) cardiac monitoring, (iii) sleep stage monitoring and (iv) clinical outcome prediction. For each category, we describe representative tasks, the corresponding biosignal datasets, and the associated challenges in data collection, labelling, and modelling.

2.2.1 Physical activity and fitness monitoring

Physical activity plays a central role in both disease prevention and functional capacity assessment. In this thesis, we focus on two representative tasks, namely human activity recognition (HAR) and cardio-respiratory fitness (CRF) prediction.

Human activity recognition task. Human activity recognition (HAR) refers to the task of identifying specific actions performed by an individual, such as walking, running, sitting, or lying down, based on measurable signals of their motion or posture. HAR is widely used in healthcare and well-being applications, enabling the detection of physical activity patterns, the assessment of mobility impairments, and the continuous monitoring of the progress of exercise or rehabilitation (Vrigkas et al., 2015). Large-scale HAR studies using built-in accelerometers in mobile devices have been used for global physical activity surveillance, revealing activity inequalities between countries based on data from more than 700,000 participants.

In general, HAR data can be collected through two main approaches. The first is video-based HAR, in which activity labels are derived from surveillance or camera systems (Lin et al., 2008). While such methods can achieve high recognition accuracy, they raise significant privacy concerns, require fixed infrastructure, and are not well suited for continuous, daily, or population-scale monitoring. The second is sensor-based HAR, which uses motion sensors, most commonly IMUs embedded in smartphones, smartwatches, or dedicated wearable devices, to capture acceleration, angular velocity, and orientation data (Zhang et al., 2022b; Tang et al., 2021). This modality offers a non-intrusive and scalable way to collect activity data in both laboratory and free-living environments. Different physical activities produce different patterns in motion signals. For example, walking generates rhythmic, repeated acceleration oscillations corresponding to each step, whereas sitting results in largely constant, low-variance acceleration patterns that reflect the body being mostly stationary.

To assess the usefulness of motion-based biosignals for healthcare tasks, this thesis examines two representative datasets: the UCI HAR dataset and the 3D HAR dataset.

UCI HAR dataset. The UCI HAR (Anguita et al., 2013) dataset is one of the most widely used benchmarks in human activity recognition. It contains recordings from 30 volunteers aged between 19 and 48 years, each performing a standard protocol of six activities of daily living, categorised as standing, sitting, lying down, walking, walking upstairs, and walking downstairs. Data were collected using a waist-mounted Samsung Galaxy S II smartphone equipped with a triaxial accelerometer and a triaxial gyroscope, sampled at 50 Hz.

Each participant performed the activity protocol twice. In the first trial, the phone was fixed on the left side of the belt, while in the second, participants positioned the phone themselves. Between activities, a five-second rest period was included to facilitate repeatability and improve ground-truth labelling. Although data collection took place under controlled laboratory conditions, participants were encouraged to perform the activities naturally to make the dataset more representative of real-world movement patterns. The dataset provides pre-segmented windows of sensor readings, each labelled with the corresponding activity. The class distribution is relatively balanced across the six activities, and the total dataset comprises over 10,000 labelled time-series segments, each lasting 2.56 seconds (128 readings per window).

3D HAR dataset. The 3D HAR dataset (Vidulin and Krivec, 2010) contains motion-capture data from five individuals performing a set of physical activities, including walking, sitting, lying, falling, and standing up. Each participant wore four motion capture tags placed on the belt, chest, and both ankles, which recorded the 3D position data of the sensors. The original recordings were continuous sequences, each comprising approximately 6,600 time points sampled in repeated scenarios.

To prepare the dataset for analysis, the raw sequences of the four tags were merged into a single 12-channel multivariate time series and resampled at 100 ms intervals. The continuous recordings were then split into partially overlapping windows of 50 time points with a 25-point overlap, resulting in a total of 6,554 labelled sequences, each 211 time points in duration. The dataset originally contained 11 activity classes, but several closely related labels were merged to reduce ambiguity (*e.g.*, lying and lying down were combined into the category lying; sitting and sitting down were combined into sitting). The final set of labels comprises seven activity categories: walking, falling, lying, sitting, standing up, on all fours, and sitting on the ground.

Unlike UCI HAR, which is regularly sampled and collected from a single device, the 3D HAR dataset involves multiple sensors with natural synchrony, making the multivariate time series inherently irregular. This property presents additional modelling challenges, such as handling variable sampling rates and aligning signals from different body locations.

Cardiofitness prediction tasks. Cardio-respiratory fitness (CRF) is a key indicator of cardiovascular health, reflecting the ability of the circulatory and respiratory systems to supply oxygen to muscles during sustained physical activity (Mandsager et al., 2018). CRF is strongly and inversely associated with the incidence of cardiovascular disease, type 2 diabetes, certain cancers, and all-cause mortality (Han et al., 2022). Clinical evidence suggests that CRF can be a stronger predictor of mortality than many well-established risk factors such as hypertension, high cholesterol, or smoking. Incorporating CRF into risk models has been shown to significantly improve the prediction of adverse cardiovascular outcomes, enabling more accurate stratification of individuals at risk and supporting more targeted prevention and intervention strategies.

The gold-standard measure of CRF is maximal oxygen uptake (VO_2max), which is defined as the maximum rate at which an individual can consume oxygen during exercise. VO_2max is typically assessed through an exercise test to exhaustion while measuring respiratory gas exchange (Swain et al., 2014). For the test result to be considered a true maximal value, specific criteria must be met, such as a plateau in oxygen uptake, a levelling off of heart rate, and exceeding thresholds for respiratory exchange ratio. These tests require specialised laboratory equipment, trained personnel, and maximum physical effort from the participants. Due to the substantial costs, logistical demands, and potential health risks associated with exercise tests to exhaustion, conducting VO_2max assessments at scale has traditionally been impractical. As a result, large-scale VO_2max testing has remained inaccessible for most population cohorts, thus limiting the feasibility of population-level CRF surveillance.

Recent advances in wearable technology have opened the possibility of estimating VO_2max from free-living physiological and motion data (Lindsay et al., 2019a), allowing scalable fitness assessment without exhaustive laboratory testing. In this thesis, we evaluate models for CRF estimation using two datasets derived from large-scale population studies, specifically the Fenland Study and the Biobank Validation Study (BBVS).

Fenland dataset. The Fenland Study (Lindsay et al., 2019a) is a large-scale, ongoing population-based cohort study involving 12,435 men and women who were aged 35–65 years at the time of recruitment. Participants attended a baseline clinic visit, after which a subset of 2,100 individuals were invited to wear two devices for seven consecutive days, specifically a combined heart rate and movement chest sensor, and a wrist-worn accelerometer on the non-dominant wrist.

The chest-mounted sensor, which is attached via two ECG electrodes at the base of the sternum, records heart rate and uniaxial acceleration at 15-second intervals. The wrist accelerometer recorded triaxial acceleration at 60 Hz. Both devices were waterproof, allowing continuous wear during sleep, showering, and exercise. In addition, resting heart

rate (RHR) was measured in the clinic with participants lying supine and the RHR value was calculated as the average heart rate during the last three minutes of a 15-minute ECG.

During the clinic visit, participants also performed a submaximal treadmill test, which, along with the wearable measurements, was used to estimate their VO_2max and calculate physical activity energy expenditure (PAEE). The study consists of two phases: Phase I (2005–2015) and Phase II (from 2014 onward), with repeat measurements collected at least four years after the first visit. In Phase II, 2,675 participants returned for follow-up testing, including a similar week-long wearable monitoring protocol.

Biobank Validation Study (BBVS) dataset. The BBVS (Gonzales et al., 2021) dataset includes 181 participants from the Fenland Study who underwent an independent maximal exercise test in which VO_2max was directly measured via respiratory gas analysis. This subset provides high-quality ground truth for validating VO_2max prediction models trained on wearable free-living data. By combining continuous heart rate and movement recordings with directly measured VO_2max, the BBVS dataset serves as a crucial benchmark for evaluating the accuracy of fitness prediction algorithms.

2.2.2 Cardiac monitoring

Atrial fibrillation (AF) is the most common sustained cardiac arrhythmia in adults. AF arises from caused by disorganised electrical activation of the atria, which effectively disrupts effective contraction. AF is a major risk factor for stroke, thromboembolism, heart failure, and increased mortality. However, its detection is often challenging because many episodes are asymptomatic and occur intermittently (paroxysmal AF), making them easy to miss with short-term ECG recordings. While continuous monitoring using 24-hour devices improves detection, it generates large volumes of data that require expert review, creating both diagnostic and resource burdens (Hill et al., 2019). Therefore, the automated analysis of long-duration ECG signals offers a scalable solution for timely and accurate AF detection, reducing both missed diagnoses and clinical workloads. In this thesis, we investigate ML approaches to identify AF and related arrhythmias from extended ECG recordings using the MIT-BIH dataset.

MIT-BIH dataset. The MIT-BIH Atrial Fibrillation database (Moody, 1983) is a well-established benchmark for AF detection. This dataset contains 25 long-term ECG recordings, each approximately 10 hours in duration, collected from human subjects with atrial fibrillation. For each recording, two ECG leads are sampled at 250 Hz, providing high-resolution waveforms suitable for rhythm analysis. The signals are annotated based on four rhythm types: (i) atrial fibrillation. (ii) atrial flutter. (iii) atrioventricular (AV) junctional rhythm and (iv) all other rhythms. For our experiments, the continuous ECG signals are

segmented into 10-second windows (2,500 samples per lead) to create manageable time series segments for analysis. The dataset is highly imbalanced, with atrial flutter and AV junctional rhythms each representing fewer than 0.1% of all segments. This, in turn, poses significant challenges for classification tasks. This imbalance motivates the exploration of unsupervised or self-supervised representation learning approaches, which can leverage unlabelled data to learn robust features before fine-tuning using limited labelled examples.

2.2.3 Sleep stage monitoring

Sleep quality is a critical determinant of human health, influencing physical recovery, cognitive performance, and emotional well-being. Poor sleep quality is linked to increased risk of stress, anxiety, daytime fatigue, obesity, hypertension, diabetes, and neurocognitive disorders. Human sleep is typically divided into two broad categories: non-rapid eye movement (NREM) sleep—which includes light sleep (stages N1 and N2) and deep sleep (stage N3)—and rapid eye movement (REM) sleep. The transitions between these stages follow characteristic cycles over the course of the night, and disruptions to this structure are often indicative of sleep disorders.

Accurate sleep stage classification is essential for assessing sleep quality and diagnosing disorders such as insomnia, narcolepsy, or sleep apnea (Morokuma et al., 2023). Gold-standard staging uses polysomnography (PSG) in a sleep lab, recording EEG, EOG, EMG, ECG, SpO_2, and airflow. While highly accurate, PSG is complex, expensive, and intrusive, requiring specialised equipment and trained personnel, which limits its accessibility in community healthcare settings. This motivates the development of automated sleep stage classification methods that can operate with fewer sensors, enabling scalable and less obtrusive sleep monitoring.

Sleep-EDF. The Sleep-EDF database (Kemp et al., 2000) is a widely used benchmark for sleep stage classification. It contains 197 full-night PSG recordings, each including EEG, EOG, chin EMG, and event markers. The recordings are segmented into 30-second epochs and manually annotated by sleep experts according to five stages: Wake (W), N1, N2, N3 (deep sleep), and REM. In this thesis, we use a single EEG channel (Fpz–Cz) sampled at 100 Hz, focusing on minimal-sensor configurations to align with real-world wearable applications.

2.2.4 Clinical outcome prediction

The ICU is a highly resource-intensive environment dedicated to the treatment of critically ill patients. Continuous monitoring of patients in the ICU is essential for detecting early signs of deterioration and initiating timely interventions, which can significantly reduce

morbidity and mortality. Predictive models that can anticipate adverse outcomes based on routinely collected clinical measurements hold great promise for improving decision-making, guiding treatment strategies, and optimising resource allocation (Alghatani et al., 2020; Aldhoayan and Aljubran, 2023). In the ICU, vital signs such as heart rate, respiratory rate, blood pressure, oxygen saturation, and body temperature are continuously monitored by biosensors, including ECG devices, along with laboratory test results and other clinical observations. Current monitoring approaches often rely on raw data being transmitted to healthcare professionals for manual or semiautomated interpretation, although this can be time-consuming and reactive. Therefore, intelligent automated systems are needed that can analyse heterogeneous high-dimensional ICU data in real time, generate timely alerts and support proactive clinical care (Iwase et al., 2021).

A major challenge in ICU data analysis is the irregularity and incompleteness of measurements. Data are often collected from multiple devices with different sampling rates, leading to asynchronous and non-continuous time series. Effective ML models for ICU prediction tasks must therefore be capable of handling missing values, irregular sampling, and multimodal inputs.

MIMIC-III. The MIMIC-III database (Johnson et al., 2016a) is a publicly available critical care dataset that covers 53,423 adult admissions to the ICU between 2001 and 2012. It contains a rich set of features, including patient demographics, vital signs, laboratory test results, medications, clinical notes, and imaging reports. For this work, we select a subset of 21,250 patients with sufficient observations and extract 96 longitudinal real-value variables from the first 48 hours after admission to the ICU. These include vital signs, laboratory measurements, and other physiological parameters. The primary prediction task is in-hospital mortality, whether the patient dies before hospital discharge, based solely on data from the initial 48 hours. This setup mimics a realistic early warning scenario in which only partial clinical information is available.

Physionet 2012. Designed similarly to MIMIC-III, the PhysioNet 2012 Challenge dataset (Silva et al., 2012) contains 12,000 patient records in the ICU, each comprising 41 clinical variables collected during the first 48 hours after admission. These variables include vital signs, lab results, and other time series measurements, recorded at irregular intervals. The dataset was specifically curated for the 2012 PhysioNet Challenge, which focused on the prediction of in-hospital mortality using incomplete, irregular, and heterogeneous clinical data. The irregularity and sparsity of the measurements make it a challenging benchmark to evaluate robust time series modelling techniques in critical care.

2.3 Modelling biosignal time series

To build effective deep learning models for these datasets, a crucial step is to extract meaningful representations from raw biosignals and learn their correlations with target healthcare outcomes. This is commonly formulated as a supervised learning problem in which, given a collection of labelled examples, a model is trained to map input biosignal sequences to task-specific outputs such as disease risk, cognitive state, or physiological parameters. The model parameters are optimised by minimising a loss function that measures prediction error and are updated through backpropagation and gradient-based optimisation (Rumelhart et al., 1986). Although the mathematical details of this process are provided in the Appendix A.1, the key idea is that supervised learning uses labelled data to guide representation learning toward task-relevant features.

Within this supervised learning framework, the choice of network architecture plays a central role in determining whether data characteristics are captured. Biosignals are inherently sequential, exhibiting temporal dependencies and correlations across multiple physiological channels (Liang et al., 2024). Consequently, modelling biosignal time series shares many similarities with general time series modelling. Historically, this field evolved from early neural network (NN) approaches to recurrent architectures, and more recently to attention-based models (Wang et al., 2024b). In addition to recurrent designs, convolutional neural networks (CNNs) have also been successfully adapted to time series analysis, providing computationally efficient alternatives.

2.3.1 Recurrent neural networks and their variants

Early approaches to modelling time series data relied on feedforward neural networks, which operate under the assumption that input samples are independent and identically distributed (i.i.d.) (Turmon and Fine, 1994). However, biosignal time series such as ECG, EEG, and PPG data are inherently sequential, where each observation is temporally correlated with its past and future values. Flattening a time series into a single vector for processing by a feedforward layer discards temporal structure, making such models ill-suited for capturing long-range dependencies. These limitations motivated the development of recurrent neural networks (RNNs) and their variants, which explicitly model sequential dependencies and have become a cornerstone in time series representation learning.

RNNs (Sherstinsky, 2020) address this limitation by introducing a recurrent hidden state that is updated at each time step, allowing the network to retain and propagate information over time. Given an input sequence $\mathbf{x} = (x^{(1)}, \ldots, x^{(T)})$, the RNN updates its

hidden state $h^{(t)}$ and outputs $y^{(t)}$ as:

$$a^{(t)} = \mathbf{b} + \mathbf{W}_h h^{(t-1)} + \mathbf{U}x^{(t)}, \tag{2.1}$$
$$h^{(t)} = \tanh(a^{(t)}), \tag{2.2}$$
$$o^{(t)} = \mathbf{c} + \mathbf{V}h^{(t)}, \tag{2.3}$$
$$\hat{y}^{(t)} = \text{softmax}(o^{(t)}), \tag{2.4}$$

where $\mathbf{U}$, $\mathbf{W}_h$, and $\mathbf{V}$ are trainable weight matrices, and $\mathbf{b}$ and $\mathbf{c}$ are biases. By sharing parameters across time, RNNs can, in principle, capture dependencies over arbitrary sequence lengths.

Limitations and the vanishing gradient problem. In practice, training RNNs on long biosignal recordings is challenging due to the problems of vanishing and exploding gradients (Sherstinsky, 2020). This means that while RNNs can capture short-term dependencies, they often struggle with long-term dependencies, which are important in biosignals for tasks such as detecting slow-changing physiological states or rare events.

Long Short-Term Memory (LSTM) networks. The LSTM architecture (Yu et al., 2019) was introduced to address the problem of vanishing gradients by incorporating an explicit memory cell $c^{(t)}$ and a set of gating mechanisms that regulate the flow of information:

$$f^{(t)} = \sigma(\mathbf{W}_f x^{(t)} + \mathbf{U}_f h^{(t-1)} + b_f), \tag{2.5}$$
$$i^{(t)} = \sigma(\mathbf{W}_i x^{(t)} + \mathbf{U}_i h^{(t-1)} + b_i), \tag{2.6}$$
$$\tilde{c}^{(t)} = \tanh(\mathbf{W}_c x^{(t)} + \mathbf{U}_c h^{(t-1)} + b_c), \tag{2.7}$$
$$c^{(t)} = f^{(t)} \odot c^{(t-1)} + i^{(t)} \odot \tilde{c}^{(t)}, \tag{2.8}$$
$$o^{(t)} = \sigma(\mathbf{W}_o x^{(t)} + \mathbf{U}_o h^{(t-1)} + b_o), \tag{2.9}$$
$$h^{(t)} = o^{(t)} \odot \tanh(c^{(t)}), \tag{2.10}$$

where $f^{(t)}$, $i^{(t)}$, and $o^{(t)}$ are the forget, input, and output gates, respectively, and $\odot$ denotes element-wise multiplication. These gates allow the LSTM to decide what information to retain or discard, making it well suited for modelling long-range temporal dependencies in biosignal time series, such as multisecond HRV patterns in ECG or multiminute sleep stage transitions in EEG.

Gated Recurrent Units (GRUs). The Gated Recurrent Unit (Chung et al., 2014) simplifies the LSTM by merging the forget and input gates into a single update gate $z^{(t)}$

and removing the separate cell state:

$$z^{(t)} = \sigma(\mathbf{W}_z x^{(t)} + \mathbf{U}_z h^{(t-1)}), \tag{2.11}$$

$$r^{(t)} = \sigma(\mathbf{W}_r x^{(t)} + \mathbf{U}_r h^{(t-1)}), \tag{2.12}$$

$$\tilde{h}^{(t)} = \tanh(\mathbf{W}_h x^{(t)} + \mathbf{U}_h (r^{(t)} \odot h^{(t-1)})), \tag{2.13}$$

$$h^{(t)} = (1 - z^{(t)}) \odot h^{(t-1)} + z^{(t)} \odot \tilde{h}^{(t)}. \tag{2.14}$$

GRUs often achieve performance similar to that of LSTMs with fewer parameters and are thus widely used in biosignal modelling where computational efficiency is of primary importance.

2.3.2 Transformer architectures for biosignal time series

While RNNs and their variants process sequences step by step, attention-based architectures such as the transformer (Vaswani et al., 2017) enable direct modelling of dependencies between any two time steps, regardless of their distance from each other within the sequence. Transformers rely entirely on self-attention mechanisms without recurrent or convolutional layers, which can be advantageous for long biosignal recordings where events far apart in time may still be related (Wen et al., 2022).

Self-attention mechanism. Given an input sequence represented as a matrix $\mathbf{X} \in \mathbb{R}^{T \times d}$, the self-attention mechanism first projects it into queries ($\mathbf{Q}$), keys ($\mathbf{K}$) and values ($\mathbf{V}$):

$$\mathbf{Q} = \mathbf{X}\mathbf{W}_Q, \quad \mathbf{K} = \mathbf{X}\mathbf{W}_K, \quad \mathbf{V} = \mathbf{X}\mathbf{W}_V, \tag{2.15}$$

and computes attention weights as:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}}\right)\mathbf{V}, \tag{2.16}$$

where d_k is the dimensionality of the keys.

In biosignal applications, self-attention can learn to focus on salient physiological events (*e.g.*, arrhythmia episodes) or temporally local patterns (*e.g.*, gait cycles) that may occur at irregular intervals (Siebra et al., 2024).

Advantages and adaptations. Transformers avoid the vanishing gradient problem and are highly parallelisable, making them attractive for large-scale biosignal modelling. However, standard transformers have quadratic complexity in sequence length, which can be prohibitive for high-frequency physiological signals. To address this, adaptations such as sparse attention (Tay et al., 2020), and hybrid CNN and transformer models for

ECG (Liu et al., 2023), as well as multimodal wearable data have been explored.

In general, the choice between RNN-based and transformer-based models for biosignal time series depends on task requirements, dataset size, and other computational constraints. In later sections, we will discuss how these architectures are integrated with self-supervised learning, domain generalisation, and foundation model approaches for effective and generalisable health diagnostics.

2.3.3 Convolutional neural networks for biosignal time series

Although RNNs are designed to capture sequential dependencies, convolutional neural networks (CNNs) provide an alternative approach to modelling biosignal time series, one that is both effective and computationally efficient. Originally developed for computer vision, CNNs have been successfully adapted to one-dimensional data, where convolution and pooling operations are applied along the temporal axis to extract local, shift-invariant patterns in time (Ganapathy et al., 2018).

Specifically, the reason CNNs are suitable for biosignal time series is because they can effectively capture local temporal structures that recur over time, regardless of their exact position. Biosignal recordings such as ECG, EEG, and wearable inertial sensor data often contain local temporal structures, for example, QRS complexes in ECG, event-related potentials in EEG, or repetitive motion patterns in accelerometer data. These patterns may occur at different times but have similar shapes, making them suitable for detection via convolutional filters, which scan over time to identify local salient features regardless of their absolute position (Yang et al., 2015). Unlike RNNs, which process inputs sequentially and may suffer from vanishing gradients, CNNs can process sequences in parallel and thus capture temporal dependencies within the receptive field of the convolution kernels.

Moreover, CNNs can learn task-dependent features directly from raw or minimally processed biosignals, avoiding the need for extensive hand-crafted feature engineering (Mallick and Baths, 2024). This is particularly valuable in healthcare applications, where manually designed features may fail to generalise across sensor types, populations, or recording conditions.

1D temporal convolution. For a multi-channel biosignal segment $\mathbf{X} \in \mathbb{R}^{T \times D}$ with T timesteps and D channels (*e.g.*, leads in ECG, electrodes in EEG, or axes in IMU), a 1D convolution layer applies K learnable filters $\mathbf{w}_k \in \mathbb{R}^{L \times D}$ of temporal length L:

$$(\mathbf{X} * \mathbf{w}_k)[t] = \sum_{d=1}^{D} \sum_{\ell=0}^{L-1} w_{k,\ell,d} \, X_{t+\ell,d}. \tag{2.17}$$

Each filter learns to respond to a specific temporal pattern across all channels, producing K feature maps that emphasise different aspects of the biosignal.

Pooling layers (*e.g.*, max or average pooling) are often interleaved with convolution layers to reduce temporal resolution while increasing robustness to small temporal shifts:

$$y_t = \max_{0 \leq q < Q} x_{t+q}, \tag{2.18}$$

where Q is the pooling window length.

Compared to RNNs, CNNs are typically faster to train, more parallelisable (Gehring et al., 2017), and less prone to gradient vanishing over long sequences. However, their fixed receptive field may limit their ability to capture very long-range dependencies unless the network is made sufficiently deep or else is augmented with dilated convolutions or attention mechanisms. For biosignals with multiscale temporal structure, CNNs are often combined with recurrent or attention-based layers to balance local feature extraction with long-range temporal modelling. In this thesis, we use a one-dimensional CNN and combined it with recurrent networks to model the ECG signals.

2.4 Training paradigms

In the previous sections, we introduced backbone architectures for modelling biosignal time series, including recurrent networks, convolutional networks, and attention-based models. Although these architectures provide the foundation for feature extraction, they do not address many of the challenges encountered in real-world healthcare applications by themselves. Biosignal datasets are often sparsely labelled, domain-specific, and collected under conditions that differ substantially from the intended deployment environment (Alanazi, 2022). A model trained in a conventional supervised manner on one dataset may therefore fail to generalise to new tasks, new populations, or new acquisition settings.

In contrast, human intelligence rarely learns entirely from scratch; instead, it builds on previously acquired knowledge to adapt to novel situations. This observation motivates the development of training paradigms that can: (i) learn useful representations from unlabelled data; (ii) transfer knowledge from one domain or task to another; and (iii) improve generalisation to unseen conditions.

This thesis explores such paradigms, with a particular focus on *self-supervised learning*, *domain adaptation*, and, finally, *foundation models* for biosignal time series.

2.4.1 Self-supervised learning

As discussed in Section 1.2, a key limitation in modelling biosignal time series is the scarcity of labelled data. Deep neural networks (DNNs) typically require large-scale

labelled datasets to learn high-quality, robust representations. However, in healthcare, manual annotation is costly and time consuming and often requires clinical expertise, making such large-scale labelling impractical.

Self-supervised learning (SSL) offers a promising solution by pretraining models on large amounts of unlabelled data using automatically generated labels (or *pseudo-labels*) derived directly from the data itself. Compared to traditional supervised learning, where labels are human-annotated, SSL designs *pretext tasks* in which the model learns to predict information that is either withheld or transformed from the input (Gui et al., 2024). Examples include solving jigsaw puzzles, predicting missing segments, or classifying rotations in images. By solving these auxiliary tasks, the network learns intermediate representations that can be transferred to downstream tasks with limited labelled data (Saeed et al., 2019).

While effective, hand-crafted pretext tasks can introduce task-specific bias. For example, predicting the rotation angle of an image may encourage learning orientation-related features, but not those needed for object recognition. This limitation has motivated a shift towards more generic SSL paradigms that directly optimise the structure of the representation space, rather than relying solely on domain-specific pretext tasks.

Contrastive learning. A major recent advance in SSL is contrastive learning (Jaiswal et al., 2020), which has demonstrated strong performance in the vision, language, and speech domains. The core idea is to learn *invariant* representations by pulling together embeddings of different augmented views of the same sample (positive pairs) while pushing apart embeddings of different samples (negative pairs). Formally, given a batch of samples, contrastive learning applies domain-specific augmentations to generate multiple views, then optimises a similarity objective (*e.g.*, InfoNCE loss), one that maximises agreement between positives and minimises agreement with negatives.

Using data augmentations to define positive and negative pairs, contrastive learning eliminates the need for manual labels and encourages the model to learn features that are stable under realistic variations in the data (Huynh et al., 2022; Saeed et al., 2020). Recently, contrastive methods have been applied to health-related domains, for example, ECG data for the detection of arrhythmias and stress recognition (Kiyasseh et al., 2020), and EEG data for the scoring of sleep and the classification of cognitive state (Mohsenvand et al., 2020). These approaches consistently outperform noncontrastive baselines, underscoring the potential of contrastive SSL for biosignals.

Research gap. Although contrastive self-supervised methods consistently outperform noncontrastive baselines, existing SSL frameworks are poorly aligned with the structural properties of biosignal time series. In this dissertation, we summarise the limitations of previous work in the following directions, carefully considering the nature of biosignals.

- *Temporal dependencies and non-stationarity:* Many biosignals have long-range temporal dependencies, statistical properties that vary over time, or physiological rhythms that are crucial for interpretation. Ignoring these may lead to suboptimal representations. Standard contrastive pipelines, which assume that random augmentations suffice, may fail to capture such temporal complexity.

- *False negatives due to temporal similarity:* In time series, two segments from different recordings may share highly similar temporal patterns (*e.g.*, similar heart rate rhythms of different individuals). Existing methods often treat such pairs as negatives, and thus the model may incorrectly divide semantically similar samples, thereby affecting the quality of the representation for the prediction of downstream healthcare.

- *Data missingness and irregularity:* In real-world settings, biosignal time series are often irregular or exhibit missingness due to uneven sampling, sensor dropouts, or device-specific noise. However, existing SSL methods typically assume regularly sampled and clean data, which can lead to performance and generalisation degradation.

These limitations reveal a clear research gap: there is currently no SSL framework explicitly designed to model temporal dependencies, mitigate temporally induced false negatives, and handle irregular or missing data in biosignals. This gap motivates the developments in Chapter 3 and Chapter 5. In Chapter 3, we investigate how to incorporate biosignal-specific characteristics into contrastive learning to better capture temporal and semantic structure and to alleviate challenges arising from limited unlabelled data. Furthermore, we introduce a general-purpose framework to address the more complex problem of data missingness in Chapter 5.

2.4.2 Transfer learning and domain adaptation

In real-world healthcare applications, biosignal datasets are often small, domain-specific, and collected under controlled conditions that differ from the eventual deployment environment. For example, an ECG model trained on hospital-grade devices may fail when applied to wearable recordings collected under free-living conditions. Likewise, an activity recognition model trained on one population may not generalise to another with different demographics or movement patterns. This *domain shift*, a mismatch between the training (source) and testing (target) data distributions, can significantly degrade model performance (Guo et al., 2022).

Transfer learning. Transfer learning aims to leverage the knowledge acquired in *source domain*, which often contains abundant labelled or unlabelled data, to improve performance

in a *target domain* where labelled data are scarce (Yang et al., 2020). The term *transfer* encompasses methods for preserving and reusing previously learnt information, possibly in a different, but related domain. In terms of modern deep learning, this often involves reusing the weights (representations or embeddings) of a pretrained network, either as a fixed feature extractor for a simple downstream model (*e.g.*, logistic regression) or via fine-tuning, where the pretrained parameters are selectively adjusted—either partially or fully—to better specialise the model for the target task.

In the context of biosignals, feature extractors can be pretrained in either a supervised manner (*e.g.*, on large clinical datasets with ground-truth labels) or a self-supervised manner (*e.g.*, via SSL tasks on unlabelled free-living data). Self-supervised pretraining naturally complements transfer learning, as it enables models to learn general-purpose representations without requiring costly annotations (Zhao et al., 2024).

While transfer learning improves generalisability and accelerates training between two domains, it often assumes that the source and target domains are closely related. Specifically, a transferred model may perform poorly (Farahani et al., 2021) if there is a significant distribution shift between the source and target domains, a common scenario in healthcare (Chen et al., 2023b). For example, an ECG classifier trained in clean hospital recordings may face severe accuracy drops when applied to noisy wearable data collected under free-living conditions. In practice, biosignals collected in different environments (*e.g.*, hospitals vs. wearables) or populations (*e.g.*, young vs. elderly) may differ substantially, leading to negative transfer, wherein pretrained knowledge hinders rather than helps target performance. Moreover, fine-tuned models may overfit to small target datasets and yet remain brittle when faced with new, unseen domains.

Domain adaptation. Domain adaptation (DA) tackles this challenge by explicitly reducing the discrepancy arising between the source and target feature distributions during training, allowing models trained on the source domain to better generalise to the target domain - even when the target domain has no labels (unsupervised DA) (Ganin and Lempitsky, 2015) or only a few labels (semi-supervised DA) (Saito et al., 2019).

Two major approaches have shown particular promise, including discrepancy-based and adversarial-based methods. Discrepancy-based methods reduce the statistical distance between the representations of the source and the target arising, often by minimising measures such as the maximum mean discrepancy (MMD) or correlation alignment (Yan et al., 2017; Sun and Saenko, 2016). By enforcing similar feature distributions, these methods encourage the source-trained classifier to perform consistently on target data (Du et al., 2021)

Adversarial-based methods, inspired by generative adversarial networks (GANs), train a domain discriminator to distinguish between the source and target features while

the feature extractor is optimised to fool the discriminator (Tzeng et al., 2017). The resulting representation is predictive for the main task, yet invariant to the domain label. This approach has been applied to adapt ECG arrhythmia classifiers across acquisition devices (Niu et al., 2020) and to transfer human activity recognition models across datasets collected in different environments (Sanabria et al., 2021). Its flexibility allows integration with temporal architectures such as RNNs and transformers, making it particularly suitable for complex, multi-modal biosignals.

Research gap. Although transfer learning and domain adaptation have enabled knowledge transfer across visual and textual domains, their assumptions often break down in biosignal healthcare applications. Existing DA algorithms, including both discrepancy- and adversarial-based methods, typically rely on large, clean, and label-aligned datasets, whereas real-world healthcare deployment involves far larger and more heterogeneous distribution gaps. Many methods also assume that the source and target domains correspond to the same healthcare tasks and are relatively clean, yet biosignal domains frequently differ in sampling rate, sensor modality, and missingness patterns across applications. Furthermore, they often assume shared label distributions between domains, which is rarely true in real deployment. Consequently, conventional DA approaches perform well under controlled domain shifts but fail to generalise under large, heterogeneous, and multi-factor domain discrepancies.

Therefore, we build on transfer learning and domain adaptation, but go beyond their standard formulations to address biosignal-specific challenges. In particular, we explore how domain adaptation can be combined with advanced representation learning frameworks to capture temporal dependencies, handle label distribution mismatches, and remain robust to large-scale shifts between controlled laboratory datasets and noisy and irregular free-living biosignals, as shown in discussed Chapter 4 and Chapter 5. By jointly integrating domain alignment with our proposed representation learning frameworks, we aim not only to enhance the robustness and generalisability of healthcare models under realistic deployment conditions, but also to ensure that these models remain stable across diverse sensor modalities, sampling regimes, and label mismatch encountered in real-world biosignal applications.

2.4.3 Foundation model

The previous sections introduced self-supervised learning and domain adaptation as two powerful paradigms for learning from unlabelled data and coping with distribution shift. However, these techniques are typically applied *per task* and *per dataset*, given that a model is pretrained for a single objective, adapted to a single setting, and then re-trained again when the downstream task or domain changes (Bommasani et al., 2021). This

workflow can be brittle in healthcare, where tasks, devices, and populations vary widely and labels are scarce. Motivated by recent advances in large-scale pretraining, *foundation models* offer a complementary direction in which we train a single, reusable model on massive, heterogeneous datasets, and then adapt it to often with minimal labelled data and to many downstream tasks and domains.

Following the general notion popularised in vision and language (Achiam et al., 2023; Wang et al., 2023a), we use the term "foundation model" to denote a model trained on a scale in diverse data with generic self-supervised objectives, such that its learnt representation is widely reusable across tasks, datasets, and even modalities. In practice, the model serves as a *backbone* whose parameters are: (i) frozen and probed with simple heads, (ii) lightly adapted with parameter-efficient methods (*e.g.*, adapters or LoRA), or (iii) selectively fine-tuned for specific applications. Compared to narrow SSL pretraining, a foundation model emphasises data diversity, task diversity, and robustness that are distribution shifts encountered in deployment.

Recently, several general-purpose time series FMs have emerged with strong zero-shot or few-shot forecasting ability. For example, Chronos (Ansari et al., 2024) treats time series as a "language" by tokenising values and training language model families, with open-source releases demonstrating fast probabilistic zero-shot forecasts and downstream transfer. Moment (Goswami et al., 2024), a family of open-source foundation models for general-purpose time series analysis, has been pretrained using multiple general time series datasets ranging from electricity to weather.

Located in the field of healthcare, recent efforts have boosted pretraining directly on physiological and behavioural signals from wearables (Narayanswamy et al., 2024; Yuan et al., 2022). For example, large-scale FM trained on consumer PPG or ECG data have shown improved downstream task performance and the ability to encode demographic and health attributes - even when presented with scarce labels (Abbaspourazad et al., 2024). Other studies explore *scaling laws* and adaptation strategies for wearable FMs, demonstrating sample-efficient transfer across dozens of health-related tasks. Industry reports further emphasise the feasibility of training FMs on consumer health data at scale, highlighting growing commercial interest, increasing availability of large free-living datasets, and emerging evidence that such models can generalise to diverse real-world wellness and health applications.

Research gap. Despite these advances, existing time series FMs face several challenges in regard to biosignal applications. First, most are optimised for *forecasting* and trained primarily on heterogeneous, nonphysiological data, thus overlooking biosignal-specific challenges such as irregular sampling, missingness, and multimodal fusion. Second, while wearable FMs reduce this gap, they often (i) focus on a single modality (*e.g.*, PPG or

ECG)(McKeen et al., 2024; Jiang et al., 2024); (ii) limit reproducibility due to privacy concerns or closed-source implementations; and (iii) report limited evidence of cross-dataset, cross-device or cross-population generalisation(Mathew et al., 2024). Finally, even large-scale FMs can overfit to device-specific artefacts or population biases, raising questions about their robustness in deployment. These limitations motivate a biosignal-focused FM that (i) is robust to missing and irregular data; (ii) generalises across domains (device, site, population); and (iii) supports cross-modal inputs.

This dissertation investigates a *biosignal foundation model* that has been trained in heterogeneous physiological and wearable datasets, with objectives and training protocols designed to handle missing/irregular sampling and cross-modal inputs, while incorporating domain-robust training for cross-dataset/device transfer. In Chapter 5, we propose the general-purpose framework and evaluate the resulting backbone in cross-dataset settings to test whether a single pretrained model can effectively adapt to diverse healthcare tasks with missingness and real-world deployment scenarios.

2.5 Performance evaluation metrics

Throughout the studies in this thesis, we use several evaluation metrics to demonstrate the performance of our approaches. For classification tasks, we report key metrics including accuracy (ACC), area under the precision-recall curve (AUPRC), F1 score, and recall. For regression tasks, we report the mean squared error (MSE) and mean absolute error (MAE). For health-related prediction tasks, we additionally calculate the coefficient of determination (R^2) and the Pearson correlation to quantify predictive performance.

Using this comprehensive set of evaluation metrics, we aim to provide a thorough analysis of our approaches, emphasising their precision and reliability in healthcare diagnostic tasks. Each metric is defined formally as follows.

Accuracy (ACC). Accuracy measures the proportion of correctly classified samples:

$$\text{ACC} = \frac{TP + TN}{TP + TN + FP + FN}, \tag{2.19}$$

where TP, TN, FP, and FN denote true positives, true negatives, false positives, and false negatives.

Recall. Recall measures the fraction of true positive samples that are correctly identified:

$$\text{Recall} = \frac{TP}{TP + FN}. \tag{2.20}$$

F1-score. The F1-score is the harmonic mean of precision and recall:

$$\mathrm{F1} = 2 \cdot \frac{\mathrm{Precision} \cdot \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}}. \tag{2.21}$$

Area under the precision–recall curve (AUPRC). AUPRC summarises the trade-off between precision and recall across thresholds. It is defined as the integral of the precision–recall curve:

$$\mathrm{AUPRC} = \int_0^1 \mathrm{Precision}(r)\, dr, \tag{2.22}$$

where r denotes recall. Higher values indicate better discrimination ability in imbalanced classification settings.

Mean squared error (MSE). For regression tasks, MSE quantifies the average squared difference between predicted and true values:

$$\mathrm{MSE} = \frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2, \tag{2.23}$$

where y_i is the ground-truth value and $\hat{y}_i$ is the prediction for sample i.

Mean absolute error (MAE). MAE measures the average magnitude of absolute prediction errors:

$$\mathrm{MAE} = \frac{1}{N} \sum_{i=1}^{N} |y_i - \hat{y}_i|. \tag{2.24}$$

Coefficient of determination (R^2). R^2 measures the proportion of variance in ground-truth values that is explained by the predictions:

$$R^2 = 1 - \frac{\sum_{i=1}^{N} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{N} (y_i - \bar{y})^2}, \tag{2.25}$$

where $\bar{y}$ is the mean of the ground truth values.

Pearson correlation. Pearson's correlation measures the linear relationship between predicted and ground-truth values.

$$\rho(y, \hat{y}) = \frac{\sum_{i=1}^{N} (y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum_{i=1}^{N} (y_i - \bar{y})^2} \sqrt{\sum_{i=1}^{N} (\hat{y}_i - \bar{\hat{y}})^2}}, \tag{2.26}$$

where $\bar{y}$ and $\bar{\hat{y}}$ denote the mean of the ground truth and predicted values, respectively.

Table 2.1: Research gaps in biosignal time-series modelling and how this thesis addresses them. The table highlights three challenges existing literature suffers, including complex unlabelled biosignal data, cross-domain distribution shifts, and the absence of general-purpose foundation models, and summarises how Chapters 3–5 each propose targeted methods to overcome these gaps.

Challenge	**Training paradigms**	**Representative methods**	**Limitations**	**How this thesis addresses it**
Complex and unlabelled biosignal data	SSL	SimCLR (Chen et al., 2020); TNC (Tonekaboni et al., 2021)	Assume regular sampling and continuous biosignal data; fail to capture biosignal characteristics such non-stationarity and temporal dynamics.	Chapter 3 introduces StatioCL to capture intrinsic biosignal characteristics. Chapter 5 extends SSL to alleviate both data complexity and lack of labels.
Cross-domain distribution shift	TF, DA	DANN (Ganin et al., 2016); MMD (Tzeng et al., 2014)	Ignore label distribution shift and struggle with heterogeneous biosignals dataset or tasks generalisation.	Chapter 4 develops fine-grained domain adaptation to alleviate label-distribution. Chapter 5 integrates domain-robust pretraining to improve diverse cross-domain generalisation.
Lack of general-purpose FMs for biosignals	Modality-specific FMs	Moment (Goswami et al., 2024); CALF (Liu et al., 2025)	FMs trained on narrow single-domain and lack robustness to complex biosignal such as irregular samplings; generalisation remains limited.	Chapter 5 proposes a general-purpose FM pretrained on unlabelled, heterogeneous biosignal datasets with irregular and missing data, enabling robust cross-domain generalisation.

2.6 Summary

This chapter has introduced the background of biosignal time series, real-world healthcare tasks, and the deep learning–based training paradigms that form the foundation of this thesis. We have highlighted the key challenges that make biosignal modelling fundamentally different from vision or language domains, including label scarcity, irregular and non-stationary signal dynamics, heterogeneous data sources, and the need to capture long-range temporal dependencies. We have further emphasised the importance of designing models that are not only accurate but also robust and generalisable when deployed in diverse real-world healthcare scenarios. Through our review of existing methods, we have identified several critical research gaps that motivate the novel approaches developed in this dissertation, as shown in Table 2.1.

First, current SSL-based frameworks, although they hold strong success in solving unlabelled data challenges, they mainly follow vision models structures and neglect biosignal-specific characteristics such as non-stationarity, temporal dependencies. These limitations hinder robust representation learning under limited labels and motivate the developments in Chapter 3, where we introduce a contrastive learning framework tailored to the characteristics of biosignal time series, enabling more effective representation learning under limited labels

Second, we have shown that existing transfer learning and domain adaptation techniques struggle to cope with the distribution shifts encountered in biosignal healthcare settings, where differences in such as sensor modality, sampling rate, population demographics, label distributions, and environmental noise are common. This motivates the work in Chapter 4, which develops fine-grained domain adaptation strategies that explicitly model label-distribution mismatch and temporal heterogeneity, enabling more reliable deployment across source and target domains.

Finally, although existing FMs demonstrate generalisation capability across tasks, most are time-series or modality-specific models trained on narrow, non-physiological, or single-domain datasets. As a result, they offer limited robustness to irregular sampling, heterogeneous biosignal datasets, and cross-task variation. This gap motivates the developments in Chapter 5, where we propose a biosignal-focused foundation model trained across diverse datasets with irregular-aware masking, modality-flexible encoders, and domain-robust pretraining.

Chapter 3

StatioCL: Contrastive Learning for Time Series via Non-stationary and Temporal Contrast

The use of deep learning to model biosignal time series holds significant promise, as demonstrated by the recent literature. However, building effective and high-performing model pipelines for diverse health prediction and diagnostic settings remains a key challenge. As introduced in Section 1.2, the widespread availability of biosignal data, due to the proliferation of wearable devices and biosensors, has not been matched by equally accessible curated datasets with reliable medical annotations (Abbaspourazad et al., 2024). This scarcity of labelled data limits the development of robust and accurate models, highlighting the urgent need for unsupervised learning approaches.

As discussed in Section 2.4, self-supervised learning, especially contrastive learning (CL), has become a leading method in unsupervised learning. CL aims to learn useful representations by distinguishing between similar and dissimilar data samples, without relying on manual labels (Arora et al., 2019; He et al., 2020). It has shown great success in domains such as vision, text, and speech. However, its application to biosignal time series remains relatively underexplored. Existing CL methods for biosignals are typically CL for time series data since biosignals exhibit similar temporal characteristics as general time series data. In detail, these methods mainly adapt data augmentation techniques from vision-related approaches (Verma et al., 2021; Robinson et al., 2021; Khosla et al., 2020; Park et al., 2022), or employ domain-specific knowledge (Zhang et al., 2022d) and temporal pattern understanding (Tonekaboni et al., 2021; Eldele et al., 2021) to enhance representation learning. However, these approaches often fail to comprehensively capture the intrinsic similarities and characteristics, such as temporal dependencies and non-stationarity, which are fundamentally different from the assumptions in other modalities. As a result, current CL approaches may yield suboptimal representations and limited

performance and data efficiency in downstream tasks.

Notably, such limited performance mainly comes from the current methods generally adhere to the principle (Chen et al., 2020), that is creating positive pairs from augmented versions of the same sample and negative pairs from randomly selected, distinct samples. Although positive pairs created from two augmented views of the same sample are generally reliable, the construction of negative pairs follows a much simpler mechanism. In standard contrastive learning, negative examples are typically sampled *randomly* from the rest of the minibatch, under the assumption that samples drawn from different instances correspond to different semantic classes. However, this mechanism becomes problematic for biosignal time series. Random sampling can inadvertently select segments that originate from the same underlying physiological state. As a result, *false negative pairs* (FNPs) arise when two segments that actually share similar semantics are mistakenly treated as dissimilar. Furthermore, in the context of time series data, this mechanism also overlooks the similarities inherent in time series signal patterns, particularly in cases where temporally proximate segments exhibit similar characteristics. These mislabelled negatives hinder the effectiveness of representation learning and may lead to degraded downstream performance.

Therefore, effective avoiding of these FNPs is crucial for the development of biosignal representation models. Firstly, failing to recognise them as similar segments and arbitrarily labelling them as negative pairs may mislead models, leading to suboptimal performance in learning common patterns (Huynh et al., 2022; Sun et al., 2023). This, in turn, adversely affects the performance of downstream time series classification. Furthermore, eliminating FNPs may decrease model confusion and errors during pre-training. Given that representative information is captured during the pretraining phase, a reduced amount of data is required to fine-tune the model for downstream tasks. This aspect is of greater significance in light of the scarcity of labelled data in real-world scenarios. Consequently, the elimination of FNPs can contribute not only to improved accuracy but also to increased efficiency.

As such, we first identify two distinct types of FNPs to highlight these oversights based on non-stationarity, which are correlated with semantic patterns, and temporal dependencies, both of which are particularly pertinent to the unique nature of time series. The first type, termed *semantic false negative pairs*, occurs when pairs of the same class are mistakenly identified as dissimilar. This arises because standard CL uses a cosine similarity–based contrastive objective, where negative pairs are sampled purely at random and assumed to represent different semantic classes. However, in many biosignal datasets this assumption does not hold. For example, in human activity recognition, two segments of IMU data representing the same activity of "walking" might erroneously be considered a negative pair in CL, as shown in Figure 3.1(a). The second type, termed *temporal false negative pairs*, is unique to time series data and stems from the inherent temporal

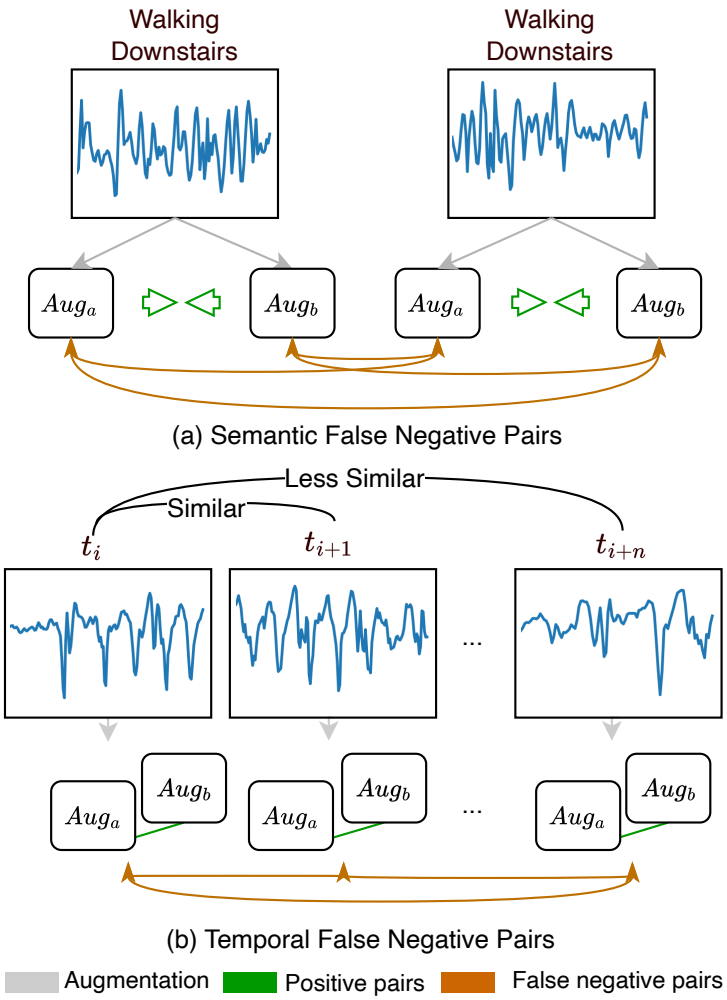


Figure 3.1: **Identification of false negative pairs** in biosignal time series contrastive learning due to random selection: (a) **Semantic FNPs**, occurring when the selection process ignores the similarity between labels and (b) **Temporal FNPs**, resulting from neglecting the similarity in terms of temporal proximity.

continuity of biosignals. These are pairs that exhibit similar patterns and are temporally proximate. Traditional CL methods might neglect this temporal continuity and incorrectly treat adjacent segments as negative pairs. As a result, two neighbouring segments with similar local dynamics (as shown in Figure 3.1(b)) may be incorrectly pushed apart and recognised as dissimilar segments. Avoiding these FNPs in time series requires a nuanced understanding of both semantic content and temporal structure. Although recent approaches have attempted to address this challenge using hierarchical clustering to refine pair construction, they often face limitations in terms of efficiency and their ability to comprehensively capture the intrinsic characteristics of time series data (Meng et al., 2023).

To address these challenges, this chapter proposes a novel CL framework named *StatioCL*[1], a specific framework designed to mitigate semantic and temporal FNPs. In particular, by effectively addressing the challenges of misleading model training caused by FNPs in time series data, StatioCL aims to improve the performance of time series classification tasks, where accurately capturing the semantic and temporal characteristics of the data is crucial for learning the common patterns between time series segments. Specifically, StatioCL employs two key strategies that harness the intrinsic characteristics of time series data: non-stationarity and temporal dependency. Non-stationarity refers to the phenomenon where the statistical properties of time series segments change in nonlinear ways, often correlating with underlying patterns. For example, in epilepsy classification, seizure segments show strong non-stationarity, while non-seizure segments are relatively stationary. By performing a statistical test to determine stationarity before training, we can obtain crucial semantic insights to help construct more accurate negative pairs. StatioCL leverages this and proposes the non-stationary contrast approach by identifying pairs with differing non-stationarity states as negative to alleviate *semantic false negative pairs*. Furthermore, to alleviate the issue of *temporal false negative pairs*, StatioCL introduces a temporal module that prioritises temporal proximity as a key factor in evaluating the similarity between segments. We propose a method to re-weight negative pairs based on their temporal distance from the anchor sample, assigning lower weights to closer segments. More specifically, the reweighting mechanism is applied within a defined proximity range, allowing for a more refined approach to negative pair selection.

This chapter makes the following contributions:

- We systematically define and differentiate for the first time two types of FNPs in biosignal time series data: semantic and temporal FNPs.
- Recognising the significant correlation between non-stationarity and temporal dependencies with regard to underlying semantics, we develop two distinct strategies

[1]The work presented in this chapter was published as a full research paper at CIKM 2024 (Wu et al., 2024).

within StatioCL to reduce FNPs and enhance representation learning.

- We conducted extensive experiments to evaluate StatioCL in four benchmark datasets of real-world time series, covering various time series modalities, including wearable and clinical biosignals. The results show an average reduction in FNPs by 19.2% and a relative improvement of 2.9% in recall for downstream classification tasks.
- Further analyses reveal intriguing properties of StatioCL concerning data efficiency and the robustness of label scarcity. In particular, when only 10% fine-tuning data is used, StatioCL demonstrates an average improvement of 3.1% is better than other CL methods and 5.4% is improved over traditional supervised methods across all data sets.

3.1 Related work

In Section 2.4, we generally discuss that, by learning from unlabelled data, CL can alleviate the need for extensive labelled datasets. This is particularly beneficial for biosignal time series, where labelled data can be scarce or difficult to acquire in healthcare settings. This section introduces how and whether existing CL methods can adequately model general time series and deal with false negative pairs.

3.1.1 Contrastive learning for time series data

Current CL methodologies for time series data can be broadly categorised into three primary streams. The first stream focusses on the development of advanced data augmentation to improve the robustness of the model (Yue et al., 2021; Zhang et al., 2022d; Yang et al., 2023; Shi et al., 2021; Luo et al., 2023). Although these augmentation techniques help generate diverse and informative positive pairs, they neglect the intrinsic characteristics of time series data, and some rely on meta-learning, which reduces the method's efficiency. Another set of strategies enhances the structure of contrastive pairs beyond the segment level, incorporating subject-level considerations by leveraging inter-subject variability or considering trial-level consistency through neighbouring and non-neighbouring samples (Lan et al., 2021; Kiyasseh et al., 2020; Wang et al., 2023b). However, these two categories are mostly inherent in the approach of constructing vision-related pairs (Chen et al., 2020) and tend to ignore the label or temporal semantics that might lead to false negative pairs. A third stream exploits the intrinsic nature of time series data, such as temporal dependencies, to enhance time series representation learning. For instance, CPC and its advanced method of Temporal and Contextual Contrasting (TS-TCC) aim to learn robust representations by predicting future states based on past states (Eldele et al., 2021; van den Oord et al., 2018). Temporal Neighborhood Coding (TNC) also leverages the

temporal dependencies with the assumption that close segments (neighbour) are similar while far away (non-neighbour) are dissimilar to design contrastive models (Tonekaboni et al., 2021) and contrasting neighbouring and non-neighbouring samples. However, these approaches often neglect finer-grained temporal continuity by arbitrarily categorising close and distant segments. As a result, current methods generally overlook label semantics or fine-grained temporal dependencies inherent in the data and generate unreliable pairs. These false negative pairs thus lead to potential inaccuracies for downstream time series classification.

3.1.2 False negative pairs in CL

Very few studies have focused on addressing the challenges of FNPs. One category of methods (Grill et al., 2020; Caron et al., 2020) only infers positive pairs through both the student and teacher models and does not contrast against negative samples. Most of these advancements are mainly focused on the image domain (Huynh et al., 2022; Sun et al., 2023). In the context of time series data, one recent approach, MHCCL, targets *semantic FNPs* by employing a hierarchical clustering method and subsequently using the cluster indices as pseudo labels to mitigate the incidence of FNPs. However, the mechanism focuses solely on semantic FNPs and significantly increases computation complexity during training, making it unsuitable for real-time processing in practical applications. The unique challenges of temporal dynamics that lead to *temporal FNPs* presented by time-series data have received comparatively less attention. Although TNC can, to some extent, alleviate the *temporal false negative pair* problem by contrasting negative samples with non-neighbours positioned far from the current segment, it still relies on random selection and treats these as absolute negative pairs, which neglects the fine-grained proximity and label semantics across the time axis.

In contrast, our work seeks to fill this gap by reducing the occurrence of both types of false negative pairs by leveraging the non-stationarity and temporal dependencies inherent in time series data. Therefore, StatioCL can navigate the intricacies of time series data to achieve improved performance and efficiency.

3.2 Methods

3.2.1 Problem definition

Given an unlabelled N multivariate time series denoted by $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^{N}$, each time series $\mathbf{x}_i \in \mathbb{R}^{T \times V}$ encapsulates a sequence of observations between V distinct variables recorded at T discrete time points. We aim to design a mapping function, f_θ, parametrised by a deep neural network and trained in a novel self-supervised contrastive scheme to learn effective

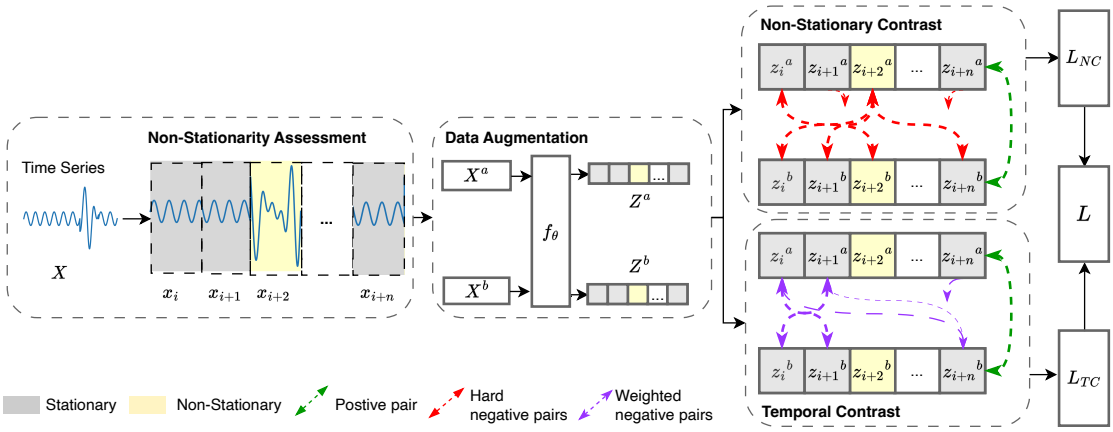


Figure 3.2: **StatioCL framework overview:** Input sequence X is first passed through **(1) Non-stationarity assessment** module to get the non-stationary state for each segment. **(2) augmentation** module will then generate two augmentation views and encode augmented views into feature space. After that, each view will be passed to **(3) Non-stationary contrast** module to construct negative pairs based on non-stationary states and reduce semantic FNPs. and **(4) Temporal contrast** module to create weighted negative pairs based on time differences and alleviate temporal FNPs. Finally, the overall loss L is calculated by combining L_{NC} and L_{TC}.

time series representations. Specifically, this scheme is distinctive due to its enhanced pair construction mechanism to address the unique challenges of false negative pairs in time series data, such as disregarding label semantics or temporal dependencies. The goal is to transform time series segments into informative representations that encapsulate inherent patterns and characteristics, thereby alleviating the challenge and facilitating a more effective and efficient application in various downstream tasks.

3.2.2 Non-stationary and temporal contrast

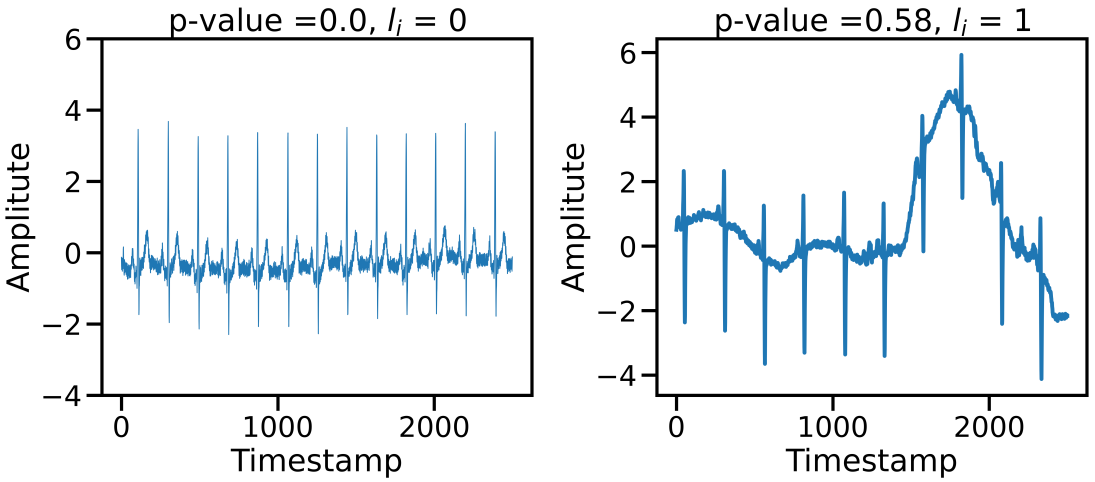
3.2.2.1 Overview

The overall contrastive learning architecture of StatioCL is shown in Figure 3.2. For each *anchor*, input segment $\mathbf{x}_i$, its non-stationary state l_i is labelled by the non-stationarity assessment module. Then, the data argumentation module will generate two augmented views using weak and strong augmentations. These augmented time series are then transformed into the latent space by the model, f_θ, to obtain the corresponding representations $\boldsymbol{z}_i{}^a = f_\theta(\boldsymbol{x}_i{}^a)$ and $\boldsymbol{z}_i{}^b = f_\theta(\boldsymbol{x}_i{}^b)$, respectively.

To mitigate the generation of false negative pairs in traditional CL methods, in the pair construction module, StatioCL employs two mechanisms that utilise the non-stationarity and temporal dependencies intrinsic within time series data to improve the quality of negative pair selection and reduce the occurrence of false negatives. The details of each module are presented below.

3.2.2.2 Non-stationarity assessment

To evaluate the non-stationary state of raw time series data, we employ the Augmented Dickey-Fuller (ADF) statistical test, a widely used method for detecting unit roots in time series data (Mushtaq, 2011). Applying this test, we compute a p-value for each segment $\mathbf{x}_i$, and set a predefined threshold depending on the characteristics of different biosignals to distinguish between stationary and nonstationary states l_i. A p-value exceeding the predefined threshold suggests the data's non-stationarity, leading us to label such segments as $l_i = 1$(non-stationary) in our framework. In contrast, segments with p values below the threshold are deemed stationary and labelled 0. This approach is exemplified in Figure 3.3, using ECG signal data, showcasing the effectiveness of our threshold-based classification in identifying *semantic false negative pairs*.



(a) Stationary ECG (b) Non-stationary ECG

Figure 3.3: **Examples of stationary and non-stationary ECG signals**. If p-value $<$ 0.01, the stationary state l_i is set to 0; otherwise, l_i is assigned a value of 1.

3.2.2.3 Data augmentation

Data augmentation is achieved by the combination of weak (jitter-and-scale) and strong (permutation-and-jitter) strategies following TS-TCC (Eldele et al., 2021). The aim is to enhance the generalisation capability of the system across different types of time series.

3.2.3 Elimination of false negative pairs

Similar to other contrastive methods, we first define positive pairs as different augmentations of the same input. Specifically, for each latent representation $\boldsymbol{z}_i$, its augmented views $\boldsymbol{z}_i{}^a$ and $\boldsymbol{z}_i{}^b$ that maintain the same stationarity state serve as the positive pair. Following

that, we use two different methods to build negative pairs in the pair construction module to alleviate both semantic and temporal false negative pairs.

3.2.3.1 Non-stationary contrast module

The non-stationary contrast module aims to address *semantic false negative pairs*. Our work constructs enhanced negative pairs by leveraging the underlying correlation between non-stationarity and label semantics in time series data. By pushing representations with distinct non-stationary characteristics away in the latent space and using a specially designed loss function, semantic FNPs can be effectively eliminated, leading to improved representations.

Specifically, this module introduces non-stationarity as an inductive bias for negative pair selection. Using the results of the Augmented Dickey-Fuller (ADF) test as prior knowledge, we assign a binary non-stationary label, l_i, to each segment $\boldsymbol{x}_i$, where $l_i = 1$ indicates a non-stationary state and $l_i = 0$ indicates a stationary state.

For the construction of contrastive negative pairs, segments with differing stationarity states are specifically chosen to form negative pairs, that is, pairing $\boldsymbol{z}_i^a$ with $\boldsymbol{z}_j^b$ where $i \neq j$ and $l_i \neq l_j$. Such pairs are designated as *hard-negative pairs*. This approach ensures robust negative pair selection, enhancing the model's ability to discriminate between distinct stationarity states and, by extension, different underlying classes. To reinforce this learning process, we introduce a tailored contrastive loss function. For a given sample x_i, the loss is computed as:

$$L_{NC}{}^{i} = -log(\frac{exp(sim(\boldsymbol{z}_i{}^a, \boldsymbol{z}_i{}^b)/\tau)}{\sum_{j=1[i\neq j, l_i \neq l_j]}^{2N} exp(sim(\boldsymbol{z}_i{}^a, \boldsymbol{z}_j{}^b))/\tau}) \tag{3.1}$$

where N denotes the batch size and τ is the temperature constant.

This customised approach effectively minimises and maximises the similarity between positive and hard-negative pairs respectively, enabling StatioCL to capture the critical relationship between non-stationarity and class distinctions, thereby eliminating *semantic false negatives pairs* and enhancing performance in downstream classification tasks.

3.2.3.2 Temporal contrast module

In StatioCL, the temporal contrast module improves the construction of negative pairs, which we term *soft-negative pairs*, to alleviate *temporal false negative temporal pairs* based on the refined assumption utilising temporal dependency. The refined assumption is that adjacent segments in time series data exhibit strong temporal dependencies over a certain period, and as this dependency diminishes over time, the similarity and relevance between these segments decrease. Then, we introduce a weighted mechanism to capture

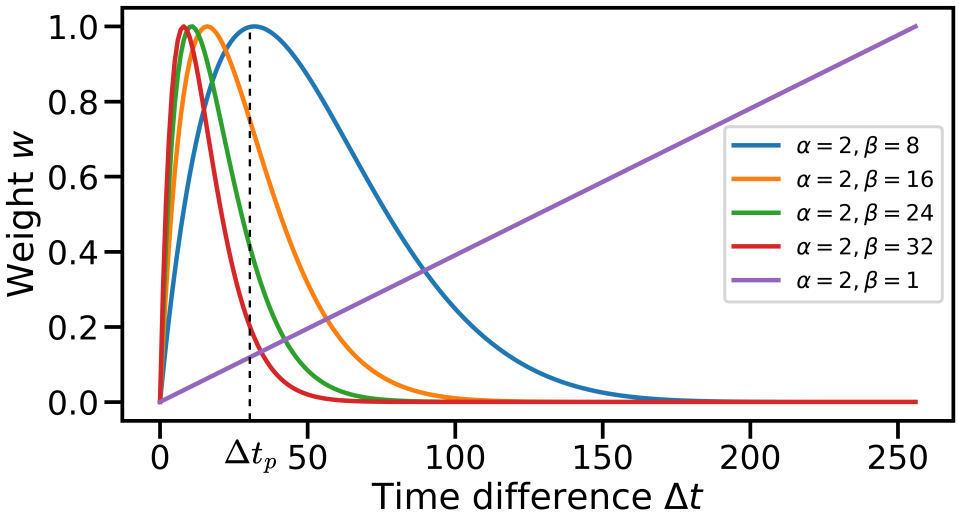


Figure 3.4: **Weight distribution for soft-negative pairs:** The histogram depicts the assigned weights to negative pairs based on their time difference, following the beta distribution.

this temporal dependency by considering the correlation between similarity and temporal proximity among negative pairs.

Specifically, these weights are parameterised using a Beta distribution to determine the degree of similarity between each latent representation $\boldsymbol{z}_i^a$ and another representation $\boldsymbol{z}_k^b$ based on their time difference, $\Delta t = |i - k|$, as illustrated in Figure 3.4. This approach is grounded in the premise that segments closer in time often exhibit higher temporal dependencies. Consequently, as the time difference increases, the segments become less similar and are more likely to form negative pairs. Initially, an increasing trend in weights is observed with respect to increasing temporal proximity. However, over a longer time span, the determination of temporal dependencies can become increasingly ambiguous. Instead of completely ignoring pairs with a longer time span, we account for scenarios where certain patterns, such as heart rate, may exhibit similar trends even over extended periods, thus making them less likely to be negative pairs. This consideration results in assigning lower weights to such pairs, leading to a decreasing trend in the latter half of the Beta distribution.

More importantly, the Beta distribution offers significant flexibility in capturing these soft negative pairs. With two parameters, α and β, controlling the weight distributions, which are learnable parameters, we can effectively approximate and optimise the temporal dynamics for different data types and applications in constructing negative pairs. For instance, with $\alpha = 2$ and $\beta = 8$ (as shown by the blue curve in Figure 3.4), the weight w_i initially increases as Δt grows, reflecting decreasing similarity. It reaches a maximum value of 1 in an optimised time range (Δt_p). As Δt extends beyond this range, reflecting a decrease in long-term temporal dependency, the influence of w_i also decreases, following a

trend that converges to zero. By varying α and β, we can learn different weight parameters, reflecting fast or slow temporal dynamics in the slope of the curves, as well as long or short temporal dependencies indicated by larger or smaller Δt_p values. Additionally, different combinations of α and β can result in linear and exponential weight distributions, thereby relaxing constraints in specific weight distribution shifts and allowing the model to learn from the data itself.

Finally, we formulate a comprehensive weighted loss function that fully enhances the aspect of temporal consistency and pair constructions. Similarly as Equation (3.1), for an input sample $\boldsymbol{x}_i$ with timestamp index i, we have the contrastive loss as follows:

$$L_{TC}{}^{i} = -log(\frac{exp(sim(\boldsymbol{z}_i{}^{a}, \boldsymbol{z}_i{}^{b})/\tau)}{\sum_{k=1[i\neq k, l_i=l_k]}^{2N} \boldsymbol{w}_i * exp(sim(\boldsymbol{z}_i{}^{a}, \boldsymbol{z}_k{}^{b}))/\tau})) \tag{3.2}$$

$$w_i = \frac{|t_i - t_k|^{\alpha-1} (1 - |t_i - t_k|)^{\beta-1}}{B(\alpha, \beta)} \tag{3.3}$$

$$B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha + \beta)} \tag{3.4}$$

where Γ is to ensure the distribution is normalised. As such, the proposed mechanism can enhance the similarity of proximate close pairs while distancing samples farther apart more softly and dynamically.

3.2.4 Overall loss

The final fine-grained contrastive loss is derived by combining the non-stationary and temporal dependency loss. This combined loss captures the signal characteristics by defining different degrees of similarity within the time series. The overall loss is defined as:

$$L = \sum_{i} [\lambda L_{NC}{}^{i} + (1 - \lambda) L_{TC}{}^{i}] \tag{3.5}$$

where λ is the scalar hyper parameter denoting the relative weight of each loss.

Therefore, with the formulation of the final loss, our approach, StatioCL, can significantly alleviate the impact of both semantic and temporal false negative pairs. This alleviation is pivotal in creating a cleaner and more accurate representation space.

3.3 Experiment setup

3.3.1 Datasets

To evaluate the effectiveness and robustness of our proposed framework, we choose popular benchmarks (as introduced in Section 2.2) that align with the experimental setup of the baseline methods and cover various domains, from daily wearable biosignals to clinical biosignals. These datasets are as follows:

- **Human Activity Recognition (HAR)**: A wearable IMU dataset with 30 subjects performing six daily activities. Signals are sampled at 50 Hz across 9 channels (Anguita et al., 2013).
- **Sleep-EDF**: Overnight PSG recordings annotated into five sleep stages (W, N1, N2, N3, REM). We use a single EEG channel following prior work (Eldele et al., 2021).
- **Epilepsy**: Short EEG recordings (23.6 s) from 500 subjects, labelled for seizure vs. non-seizure activity. We follow the binary classification setup in (Eldele et al., 2021).
- **ECG**: Long-term ECG recordings (10 h) from subjects with atrial fibrillation (AF), containing two leads sampled at 250 Hz and annotated with four rhythm classes.[2]

We divide the data into training (60%), validation (20%), and testing (20%) sets following the existing setup

Table 3.1: **Statistical summary** of biosignal time series classification datasets.

Dataset	Samples	Features	Classes	Seq. length
HAR	10299	9	6	128
Sleep-EDF	34522	1	5	3000
Epilepsy	12500	1	2	179
ECG	58766	2	4	2500

3.3.2 Baselines

We have previously discussed general self-supervised learning (SSL) frameworks in Section 2.4. In this section, we specifically compare our approach with recent SSL methods on time-series classification tasks as follows:

- **CPC** (van den Oord et al., 2018) uses autoregressive models in latent space to predict future representations, then contrasts these predictions with actual future embeddings against randomly sampled negatives.

[2] `https://physionet.org/content/afdb/1.0.0/old/`

- **SimCLR** (Chen et al., 2020) generates augmentation-based embeddings and optimises model parameters by minimising NT-Xent loss in the embedding space.
- **BYOL** (Grill et al., 2020) trains two networks simultaneously: an online network predicts the representation of the target network of the same data. The target network's weights are a moving average of the online network, thus avoiding negative pairs.
- **TNC** (Tonekaboni et al., 2021) leverages the temporal dependencies with the assumption that close segments (neighbour) are similar while non-neighbours are dissimilar to design contrastive models.
- **TS-TCC** (Eldele et al., 2021) aims to learn robust representation through a harder prediction task, *ie*, predicting the future states of the weak time series based on the context of the past strong augmented time series.
- **TF-C** (Zhang et al., 2022d) aims to improve the learning of time series representation by augmenting time series data through frequency perturbations, which demonstrates promising performance in relevant tasks.
- **MHCCL** (Meng et al., 2023) performs a hierarchical clustering contrastive learning strategy over augmented views by introducing a twofold strategy: an upward masking technique to update prototypes by removing outliers and a downward masking method for selecting contrastive pairs.

3.3.3 Implementation details

During pretraining, we employed CNNs as encoders due to their generalised capability to effectively capture distinctive features. Specifically, the network architecture includes a 3-layer 1-D CNN followed by a fully connected layer with an output dimension of 64.

To determine the appropriate levels of similarity in non-stationarity, we optimise the ADF threshold using a grid search over thresholds $\tau \in \{0.0001, 0.001, 0.01, 0.05\}$. For each dataset, we selected the value that produced the most stable non-stationary state labelling depending on the specific characteristics of each signal (measured by the variance of l_i across segments) and yielded the best downstream validation performance. For example, the IMU signal for HAR is sensitive to small body movements, resulting in noisy time series data that affect the ADF test statistics. Therefore, we use a high threshold to guarantee a reasonable l_i from the ADF test results. The chosen thresholds were: $\tau = 0.05$ for HAR, $\tau = 0.01$ for Epilepsy, $\tau = 0.001$ for Sleep-EDF, and $\tau = 0.01$ for ECG.

We conducted a grid search over thresholds $\tau \in \{0.0001, 0.001, 0.01, 0.05\}$. For each dataset, we selected the value that produced the most stable non-stationary state labelling

(measured by the variance of l_i across segments) and yielded the best downstream validation performance during pretraining. The chosen thresholds were: $\tau = 0.05$ for HAR, $\tau = 0.005$ for Epilepsy, $\tau = 0.0001$ for Sleep-EDF, and $\tau = 0.01$ for ECG.

For capturing temporal dependencies using the beta distribution, we empirically select $\alpha = 2$ and fine-tune the β value through a grid search over the set 8, 16, 24, 32. The choice of beta values depends on the nature of the data, such as whether it exhibits fast-changing temporal dynamics with short-term correlations (lower beta values) or slow-changing dynamics with long-term correlations (higher beta values). The optimised beta values are 8 for the HAR, sleep-EDF, and epilepsy datasets and 24 for the ECG waveform dataset. To control the relative weight of each loss term, we optimise $\lambda = 0.5$ for the HAR, epilepsy, and sleep-EDF datasets and $\lambda = 0.1$ for the ECG dataset.

Regarding training specifics, we set the batch size to 256 for the ECG dataset and 128 for the others due to their smaller size. We use the Adam optimiser with a learning rate of $3e^{-4}$, weight decay of $3e^{-4}$, and moment decay rates $\beta_1 = 0.9$ and $\beta_2 = 0.99$. Each model is trained for 150 epochs.

To ensure the reliability of our results, we repeated all experiments 5 times with different random seeds and reported the mean and standard deviations of the performance metrics. Finally, all models are implemented using PyTorch, and the experimental evaluations are conducted on NVIDIA A100-SXM-80GB GPUs.

3.3.4 Evaluation metrics

In line with previous studies, we employed widely recognised metrics for classification, namely, prediction accuracy (ACC) and the macro-averaged F1 (MF1) score. Due to the severe imbalance in the ECG waveform dataset, we employed the Area Under the Precision-Recall Curve (AUPRC) instead of MF1 for ECG, as it better manages the data imbalance. In addition, we incorporated recall into our evaluation to correctly identify true positive cases, considering both true positives and false negatives. This metric effectively highlights the significance of our method in reducing false negative pairs in CL, thereby enhancing the accuracy of true positive identifications and minimising instances of false negative predictions. We also employ macro average recall (MAR) to provide a balanced view of all classes, especially in datasets with varying class distributions.

3.4 Results

3.4.1 Biosignal time series classification

To demonstrate the effectiveness of our advanced method for time series classification, we first pretrain the encoder via CL and then freeze all its parameters for downstream tasks.

Table 3.2: **Comparative performance on biosignal time series classification:** Accuracy (ACC) and macro-average F1 (MF1) scores across different datasets. **Bold** indicates the best performance, while underscore indicates the second-best. StatioCL consistently outperforms all existing self-supervised learning baselines across four heterogeneous datasets and even surpasses supervised learning in most cases.

	HAR		Sleep-EDF		Epilepsy		ECG	
Methods	ACC	MF1	ACC	MF1	ACC	MF1	ACC	AUPRC
Supervised	90.1 ± 2.5	90.31 ± 2.24	83.4 ± 1.4	74.78 ± 0.86	96.7 ± 0.2	94.52 ± 0.43	94.8 ± 0.3	0.67 ± 0.01
CPC	83.9 ± 1.5	83.27 ± 1.66	82.8 ± 1.7	73.94 ± 1.75	96.6 ± 0.4	94.44 ± 0.69	68.6 ± 0.5	0.42 ± 0.01
SimCLR	81.0 ± 2.5	80.62 ± 2.31	78.9 ± 3.1	68.60 ± 2.71	93.0 ± 0.6	88.09 ± 0.97	74.6 ± 1.2	0.57 ± 0.08
BYOL	89.5 ± 0.2	89.31 ± 0.17	80.1 ± 2.2	72.34 ± 0.60	**98.1 ± 0.1**	**96.99 ± 0.15**	73.4 ± 0.9	0.55 ± 0.02
TNC	88.3 ± 0.1	88.28 ± 0.13	83.0 ± 0.9	73.44 ± 0.45	96.2 ± 0.3	94.47 ± 0.37	77.8 ± 0.8	0.55 ± 0.01
TS-TCC	89.6 ± 1.0	90.38 ± 0.29	83.0 ± 0.7	72.13 ± 1.04	96.9 ± 0.2	95.00 ± 0.24	85.0 ± 1.0	0.59 ± 0.02
TF-C	88.2 ± 0.7	88.20 ± 0.68	65.8 ± 0.7	56.27 ± 0.24	96.8 ± 0.4	94.23 ± 0.36	78.8 ± 1.0	0.56 ± 0.03
MHCCL	91.6 ± 1.1	91.77 ± 1.11	71.1 ± 0.4	61.05 ± 0.70	97.9 ± 0.5	95.44 ± 0.82	80.5 ± 0.5	0.52 ± 0.02
StatioCL	**93.1 ± 0.4**	**93.08 ± 0.29**	**83.7 ± 0.3**	**74.81 ± 0.36**	97.9 ± 0.1	95.73 ± 0.13	**87.1 ± 0.9**	**0.61 ± 0.02**

Specifically, we follow the standard linear benchmarking evaluation scheme (Chen et al., 2020), where a linear classifier is trained on top of a frozen self-supervised pretrained encoder model to assess the representations learnt on various downstream datasets. The comprehensive results are presented in Table 3.2.

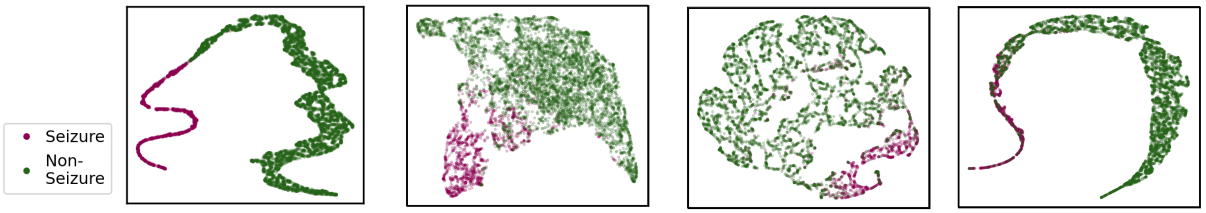
Our proposed StatioCL model showed superior performance in all datasets compared to the SOTA SSL models. In our experiments, StatioCL demonstrates an improvement in ACC ranging from 0.3% to 4.0% and an enhancement in MF1 or AUPRC ranging from 0.3% to 3.0% across five datasets compared to the most robust baselines. In particular, StatioCL even surpassed supervised learning in 4 of 5 datasets, while other SOTA SSL methods failed. This significantly underscores the efficacy of our proposed self-supervised learning framework. We also note that BYOL performs strongly on the Epilepsy dataset. This is largely because the task is a clean and well-separated binary classification problem, where the absence of negative pairs in BYOL is less detrimental than in more complex, multi-class datasets. Overall, these performance enhancements are a direct result of our innovative approach, which effectively reduces the occurrence of both semantic and temporal false negative pairs. By resolving these conflicts in representation learning, our method more accurately captures the underlying patterns in time series data, thereby significantly improving the efficacy of downstream classification tasks.

3.4.2 False negative pairs elimination discussion

We evaluated the effects of StatioCL on false negative mitigation by comparing the reduction in the portion of FNPs, according to the ground-truth labels, during the training process and compared it with SOTA CL (TS-TCC). By effectively reducing the occurrence of FNP, StatioCL not only improves the accuracy of data representation learning, but it is also expected to significantly enhance the model's ability to correctly recognise positive

Table 3.3: **Reduction in false negative pairs performance:** False negative pairs (FNPs, lower is better) and macro average recall (MAR, higher is better) for TS-TCC and StatioCL. StatioCL achieves substantially lower FNP rates, reducing them by average 19% and consistently improves MAR.

	HAR		Sleep-EDF		Epilepsy		ECG	
Method	FNP↓	MAR↑	FNP↓	MAR↑	FNP↓	MAR↑	FNP↓	MAR↑
TS-TCC	16.9%	93.1 ± 0.5	27.2%	89.6 ± 0.9	68.5%	72.1 ± 0.8	46.2%	84.2 ± 1.5
StatioCL	**3.9%**	**95.1 ± 0.3**	**18.5%**	**92.7 ± 0.9**	**22.4%**	**74.1 ± 0.4**	**26.8%**	**86.7 ± 0.8**



(a) Supervised (b) SimCLR (c) TS-TCC (d) StatioCL (Ours)

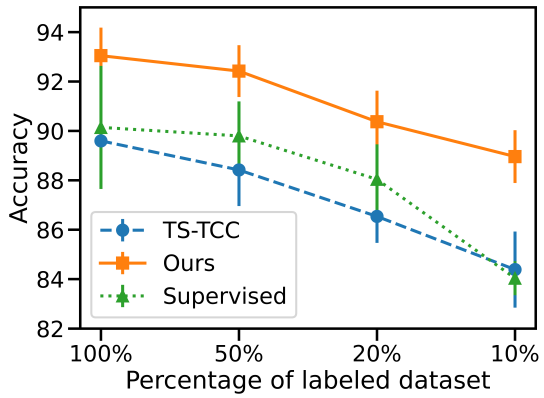
Figure 3.5: **Learned representation embedding spaces on the epilepsy training set.** StatioCL produces a cleaner, more separable latent space than prior contrastive methods, showing the benefit of reducing false negative pairs.

cases, which can be particularly evident in the recall metric. Such a capability is crucial in medical applications where the cost of missing a true positive case is high.

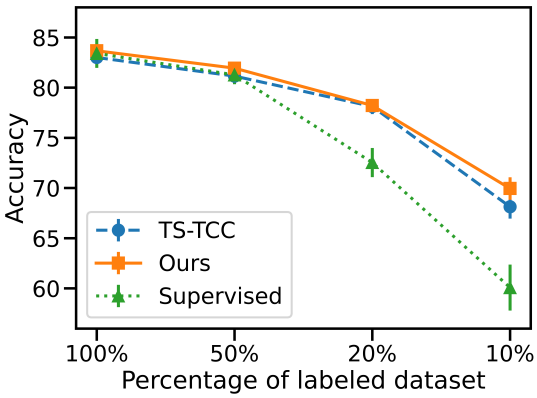
As detailed in Table 3.3, we observed a substantial reduction in the proportion of FNPs with StatioCL. Specifically, our framework achieved an average 19.2% reduction in FNPs compared to TS-TCC. This reduction directly contributed to an improved recall performance. StatioCL surpasses the SOTA baselines on an average of 2.9%, underscoring its effectiveness in accurately identifying true positive cases. Moreover, alleviating false negative pairs contributes to a more reliable embedding space. As evidenced in Figure 3.5, StatioCL achieves better class separation than traditional contrastive learning methods and captures a similar embedding shape to the supervised method. By effectively reducing these ambiguities, StatioCL enhances the clarity of class distinctions and potentially lowers the risk of misclassification, thereby underscoring the method's efficacy in accurately capturing the underlying patterns of time series data.

3.4.3 Efficiency analysis

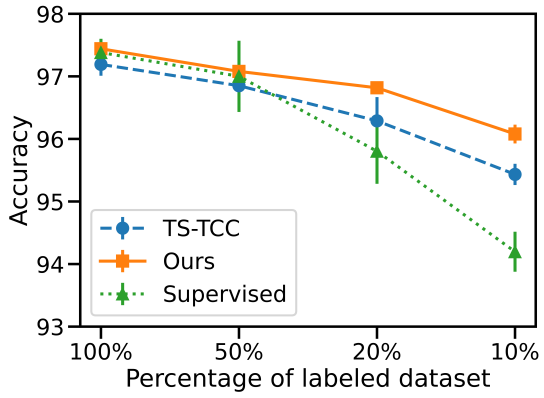
Aside from evaluating CL models on downstream datasets with a full range of labelled data, our proposed method was also assessed when labelled data is limited and less training data is available, a common occurrence in real-world scenarios due to constraints on data collection resources. This is particularly important for time series classification, where labelled data can be scarce or expensive to obtain. By varying the size of the labelled



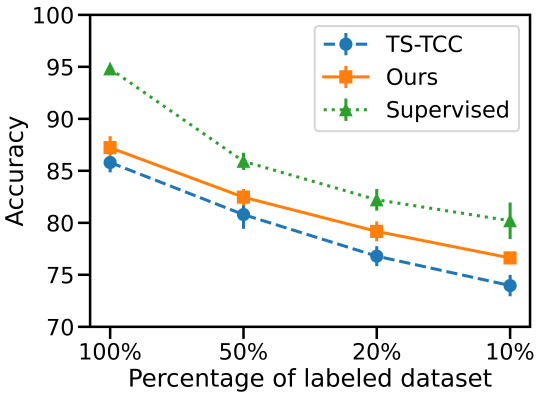
(a) HAR



(b) Sleep-EDF



(c) Epilepsy



(d) ECG

Figure 3.6: **Label efficiency analysis:** This figure shows the performance of efficiency analysis of StatioCL by varying the size of the labelled training dataset from 100% to a sparsely labelled dataset (10%). Across all datasets, StatioCL maintains higher accuracy then SOTA CL methods and exhibits a smaller performance drop as labelled data decreases, demonstrating strong robustness under limited supervision.

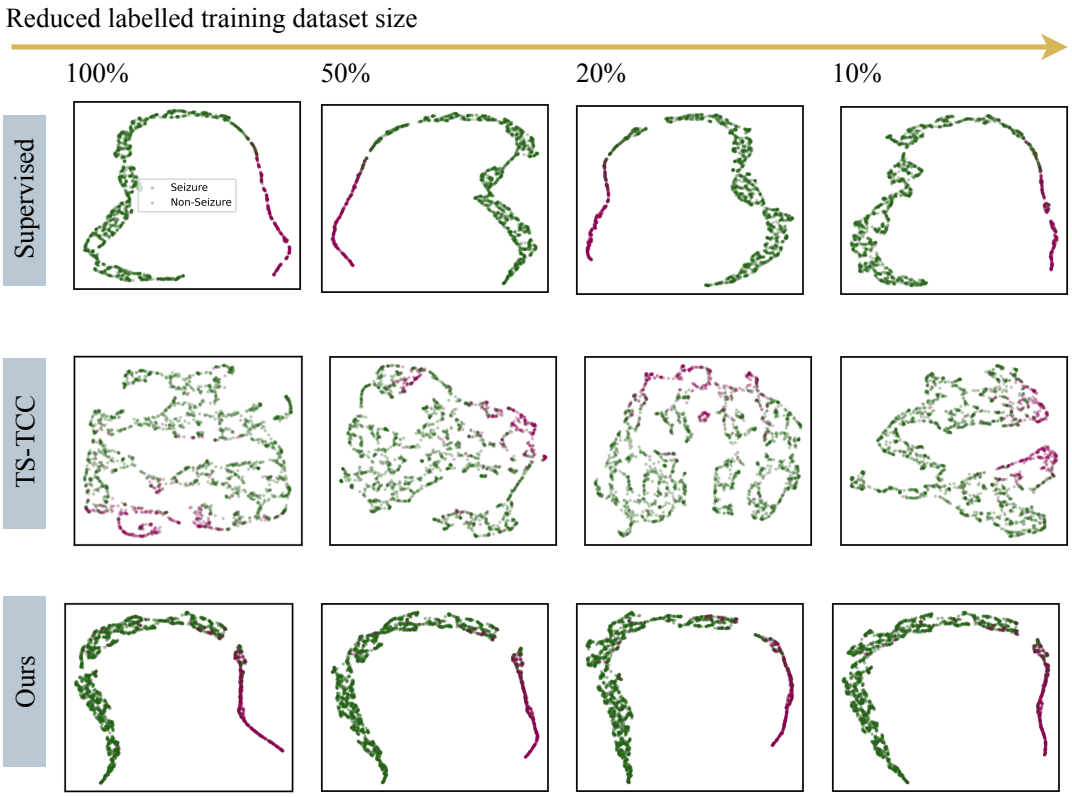


Figure 3.7: **Learned representation on the test set of epilepsy** when labelled training data size reduces from 100% to 10%. StatioCL keep producing a cleaner, more separable latent space than prior contrastive methods, showing the effectiveness and efficiency of reducing FNPs.

training dataset from a full range (100%) to a sparsely labelled dataset (10%), we further validated the efficiency and robustness of StatioCL in handling data scarcity.

In Figure 3.6, we first present a performance comparison regarding ACC for all five datasets when the labelled training dataset decreases from 100% to 10%. In particular, StatioCL consistently outperforms the SOTA TS-TCC method in all databases, showing a smaller performance drop with a data labelled with reduced availability. Specifically, TS-TCC shows an average of 9.8% relative decrease in performance across the datasets, while StatioCL only experiences an 8.2% relative decrease. Remarkably, our model still surpasses supervised learning in four out of five datasets, even when only 10% of labelled data are available. For the ECG dataset, our method outperforms all existing SSL baselines but remains slightly below the supervised model. This is likely due to the severe class imbalance in the ECG data, where supervised learning offers stronger guidance than self-supervised pretraining alone. Together, these results highlight the robustness and efficiency of StatioCL under low-data conditions. Furthermore, Figure 3.7 reveals how StatioCL maintains coherent latent space representations in the epilepsy dataset, even

Table 3.4: **Ablation study of StatioCL across four datasets.:** Removing either the temporal contrast (TC) or non-stationary contrast (NC) module consistently reduces performance, showing that both components are essential. The full StatioCL achieves the strongest and most stable performance across all tasks.

	HAR		Sleep-EDF		Epilepsy		ECG	
Methods	ACC	MF1	ACC	MF1	ACC	MF1	ACC	AUPRC
StatioCL	**93.1 ± 0.4**	**93.1 ± 0.3**	**83.7 ± 0.3**	**74.8 ± 0.4**	**97.4 ± 0.1**	**95.7 ± 0.1**	**87.0 ± 0.9**	**0.61 ± 0.02**
w/o TC	90.3 ± 0.3	90.3 ± 0.4	78.1 ± 1.2	70.4 ± 1.1	95.3 ± 3.1	94.9 ± 1.4	81.0 ± 0.3	0.60 ± 0.02
w/o NC	73.3 ± 4.8	72.3 ± 5.2	64.8 ± 2.6	54.1 ± 3.5	85.5 ± 3.0	68.8 ± 2.6	85.1 ± 0.8	0.59 ± 0.01

as the labelled data are reduced. In contrast, supervised methods and TS-TCC struggle to preserve distinct class discrimination under such constraints, leading to a decline in classification performance. These findings confirm the effectiveness of StatioCL in low-data regimes and showcase its ability to correct misleading embeddings caused by FNPs. Hence, StatioCL directly addresses the critical challenges of contrastive learning, improving model efficiency and accuracy, particularly in fine-tuning with less training and labelled data, where FNPs could significantly distort the learning process.

3.4.4 Ablation study

To evaluate each module's contribution in StatioCL, we compare the full StatioCL against two variants across all datasets (Table 3.4). Firstly, removing the temporal contrast (TC) module decreases performance, highlighting its importance in capturing fine-grained temporal dependencies and reducing temporal FNPs. Additionally, we excluded the non-stationary contrasting (NC) module, a higher decrease is observed compared to StatioCL w/o TC in 4 out of 5 datasets. This underscores NC's role in enhancing label semantics understanding and mitigating semantic FNPs.

Interestingly, for the ECG dataset, we noticed that the TC module plays a more important role than the NC module. To explore this, we analysed the correlation between non-stationarity and label semantics by examining the ratios of stationary to non-stationary segments within each class. Generally, a pronounced difference in these distributions among classes suggests a more evident correlation between non-stationarity and downstream task labels, which in turn augments the efficacy of the non-stationary contrast. Our examination revealed that in the ECG dataset, the ratios are fairly consistent across classes, with similar values of 5.71 and 6.35 for dominant classes 0 and 3, respectively, for class 0 and 6.35 for class 3. In contrast, the epilepsy dataset exhibits obvious differences: 18.1 for the seizure class versus 2.1 for the non-seizure class. Such a significant difference inherently supports the non-stationarity contrast, leading to a more significant impact of this module. These findings demonstrate the effectiveness of the non-stationary and temporal contrast modules for accurate time series representation learning.

3.5 Discussion and conclusions

In this chapter, we addressed Research Question 1, which asked how we can learn effective representations from unlabelled biosignal time series data to achieve strong performance. To this end, we introduce StatioCL, a novel contrastive learning framework for biosignal time series representation. By comprehensively capturing the inherent complexities of time series data, including non-stationarity and fine-grained temporal dependencies, StatioCL mitigates the issue of misleading training caused by both semantic and temporal FNPs. Through extensive experiments, we demonstrated that our framework achieves state-of-the-art performance on real-world biosignal benchmarks, spanning both wearable and clinical datasets. In particular, StatioCL improves the efficiency and robustness of the data when the labelled data is limited, highlighting its value for a wide range of downstream healthcare applications. These results suggest that StatioCL advances the understanding of intrinsic biosignal properties and provides a strong foundation for future work in robust representation learning.

Despite these contributions, it is important to acknowledge the limitations of this study. First, our framework relies on the ADF test as a heuristic measure to distinguish stationary and non-stationary segments for semantic state construction. Although we mitigate dataset-specific sensitivity through grid search, the ADF threshold remains a model-based statistical assumption rather than a learned, data-driven parameter. As such, the correlation between non-stationarity and label semantics, while evident in several of our datasets, is not strictly guaranteed across the full range of biosignal types. This reliance on a predefined statistical test may therefore limit robustness and generalisation when applied to unseen biosignals. Furthermore, our experiments focus on in-domain training and evaluation: pretraining and fine-tuning are conducted on the same dataset, with the distinction that pretraining uses unlabelled data. Although this setting validates the utility of self-supervision, the question of generalisation across domains and tasks remains open and is left for further exploration. To this end, the following chapters extend beyond StatioCL by exploring methods that enhance cross-domain generalisation.

Chapter 4

UDAMA: Unsupervised Domain Adaptation through Multi-discriminator Adversarial Training with Noisy Labels

In the previous chapter, we introduced StatioCL, a contrastive learning framework that learns robust biosignal representations from large-scale unlabelled data. Although this approach demonstrated strong in-domain performance, it assumed access to clean and well-annotated datasets for downstream tasks. However, in reality, healthcare datasets are often limited in size, sparsely labelled, and annotated under controlled or clinical conditions with *gold standard* labels. This setting hinders the generalisation of models to unseen cohorts and applications, as highlighted in Section 1.2.

A key challenge in this context lies in the nature of distribution shifts. Most prior work on domain adaptation has focused on changes in the feature space, such as differences in sensor hardware (Ye et al., 2024). In contrast, healthcare data are frequently affected by distribution shifts in the label space, where the definition, accuracy, or quality of labels varies depending on how they are obtained. This distinction is particularly important in biosignal modelling, yet it has been largely overlooked in the existing literature.

Collecting high-quality labels (*i.e.*, gold-standard) for healthcare applications may require a great deal of effort and can be particularly time consuming. For example, developing a precise epileptic seizure diagnosis model requires electroencephalography (EEG) during ambulatory screening (Shoeibi et al., 2020) to detect accurate brain function and a diagnosis from physicians or neurologists to label the occurrence of seizures. Consequently, most existing datasets tend to be small-scale, leading to poor performance and model generalisation (Raschka, 2018) when developing DL models for different cohorts.

In comparison, with the widespread use of mobile and wearable devices, large-scale

less accurate labels, called *silver-standard*, are available. For example, heart rate or oxygen saturation (SpO_2) can be easily gathered from smartwatches in daily life without clinical visits. However, these silver-standard labels, which are derived from a less accurate estimation scheme, often lead to predictions with lower accuracy, as they are often noisy and display distribution shifts compared to gold-standard labels (Karimi et al., 2020).

As such, gold-standard labels are crucial for the development and validation of robust clinical models. Although silver-standard labels with extensive labelling are easy to access, they usually contain noise due to less accurate collection schemes and are characterised by distribution shifts, which makes validation against gold-standard data difficult. In this chapter, we answer the following question: *Can we leverage large-scale noisy silver-standard datasets to improve deep learning model validation on gold-standard datasets for healthcare applications?*

Motivated by the challenge, we present UDAMA[1], an unsupervised domain adaptation technique with multi-discriminator adversarial training framework, tailored for healthcare label distribution shifts. Our proposed framework is inspired by adversarial-based domain adaptation methods as discussed in Chapter 3, which often contain a specific discriminator, one that categorises the source or target domains from which samples originate. However, recent work on domain adaptation still faces limitations that they mainly concentrate on large-scale gold-standard source labels to adapt to unlabelled target data, while neglecting the fact that the source domain might contain noisy silver-standard labels. Specifically, existing approaches that focus mainly on discriminating domains as a binary classification task neglect the variance of the label distribution. In contrast, UDAMA captures the fine-grained information that resides within the label distribution shifts. To achieve this, we introduce a fine-grained discriminator that attempts to discriminate the distribution of domain labels, thereby capturing domain-invariant feature representation learning. As a result, through a multidiscriminator learning scheme, UDAMA can effectively learn cross-domain representation by integrating coarse and fine domain information, resulting in promising performance in small-scale datasets.

To demonstrate the effectiveness of our framework in real-world healthcare scenarios, we focus on the task of cardiorespiratory fitness prediction (CRF), a representative real-world setting where large-scale free-living biosignals collected from wearables are available, while most existing frameworks have only been validated on small, controlled datasets. CRF is a significant predictor of cardiovascular disease (CVD) (Laukkanen et al., 2001), one of the leading causes of death worldwide (Kaptoge et al., 2019). CRF is directly measured by maximal oxygen consumption (VO_2max), which is assessed using heart rate responses to standard maximal exercises tests (*i.e.*, gold-standard). However, collecting VO_2max labels

[1]The work presented in this chapter was published as a full research paper at MLHC 2023 (Wu et al., 2023).

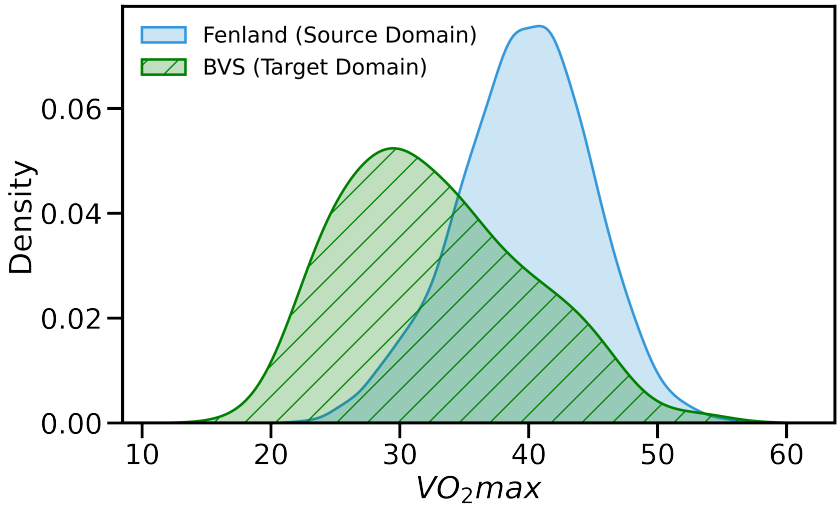


Figure 4.1: **Density distribution comparison plot:** The figure highlights a noticeable mismatch in VO_2max distributions between gold-standard target domain and silver-standard annotated source domain, motivating the need for domain adaptation.

through such tests is difficult and expensive and requires additional equipment and clinical monitoring. Alternatively, submaximal VO_2max tests (Gonzales et al., 2020b) have been proposed to capture fitness levels. Despite their potential, such measurements have been shown to provide silver-standard labels with lower accuracy and a shift in distribution compared with gold-standard VO_2max as shown in Figure 4.1.

Our experiments used two real-world datasets including a large silver, noisy labelled dataset with 12,435 participants and gold-standard labelled dataset with 181 participants to validate our model's generalisability to mitigate label distribution shift.

This chapter makes the following contributions:

- We propose a novel domain adaptation framework via multi-discriminator adversarial training, one which incorporates samples from different distributions and allows us to learn better feature representations for (often small) gold-standard datasets.

- In the CRF prediction task, gold-standard VO_2max values are hard to obtain, and silver-standard labels display domain shift problems. We are able to address the domain shift problem using UDAMA, significantly improving the prediction accuracy in the target domain. Furthermore, we stress-test our models with semi-synthetic data under various label shifts to show and test for robustness.

- Through a set of extensive experiments, we show that UDAMA achieves strong results (corr = 0.701 ± 0.032) and improves model performance up to 12.0% compared to baselines in CRF tasks. These results show that the proposed model is capable of improving fitness prediction using only large-scale silver-standard labels.

4.1 Related Work

As discussed in Section 2.4, transfer learning and domain adaptation are widely used in machine learning to address cross-domain distributional differences. This section examines in detail how existing methods alleviate such challenges. We begin with an overview of cardio-fitness estimation methods, followed by related studies that incorporate domain adaptation.

4.1.1 Cardio-fitness estimation

Numerous prediction models have recently been developed using different testing schemes (submaximal exercise or nonexercise tests) and a variety of ML methods (Jensen et al., 2021; Abut et al., 2016) to substitute direct measurements of VO_2max, the direct indicator of CRF. Specifically, with the help of modern wearable technology, which can track physical activity, resting heart rate (RHR), and other biosignals, various silver-standard methods have emerged for more convenient VO_2max calculation without maximal exercise testing (Plasqui and Westerterp, 2006; Esco et al., 2011; Shcherbina et al., 2017; Henriksen et al., 2018; Perez-Pozuelo et al., 2021). For example, deep learning models are utilised to predict fitness prediction converting raw wearable sensor data (Spathis et al., 2022). Nonetheless, these approaches overlook the impact of silver-standard labels generated from imprecise testing schemes, which can result in diminished model performance and untrustworthy fitness forecasts. Recently, a few works have been proposed that apply ML models to CRF prediction using gold-standard VO_2 max labels (Abut and Akay, 2015). However, they predominantly test only on small cohorts, which might lead to poor generalisation performance. In this chapter, we propose a novel DL method, one that aims to alleviate the distribution difference between imprecise silver-standard and gold-standard data for better fitness prediction (Results are discussed in Section 4.5).

4.1.2 Domain adaptation

A domain combines the input population with the output following a certain probability distribution (Kouw, 2018). DA is one of the state-of-the-art solutions (Hoffman et al., 2018) to learn information from an abundant labelled source domain and apply it to a new and unseen target domain with a different distribution. DA trains a feature extractor to learn the shared information between domains, so the model can generalise to new settings and thus mitigate domain change (Kouw, 2018). Discrepancy-based and adversarial-based DA approaches are effective among extant DA methods.

4.1.2.1 Discrepancy-based method

For discrepancy-based approaches, the network typically shares or reuses the initial layers between the source and target domains and aims to reduce feature space divergence (Wang and Deng, 2018). The maximum mean discrepancy (MMD) (Gretton et al., 2012) is an effective distance metric that measures the distribution divergence between the mean embeddings of two distributions in the reproducing kernel Hilbert space (RKHS) to minimise the difference between two distributions. Many methods (Tzeng et al., 2014; Long et al., 2015) utilise the MMD metric within the network to learn domain-invariant and discriminative representations. On the other hand, the correlation alignment (CORAL) (Sun and Saenko, 2016) method is proposed to align the second-order statistics of the source and target distribution with a linear transformation. Deep-CORAL (Sun and Saenko, 2016) is an extension of CORAL that can can train a non-linear transformation to align the correlations of the representation embedding within DNNs from the source domain to the destination domain. Some works, which seek to minimise the empirical Wasserstein distance arising between source and target feature representations, also showed good results in domain-invariant representation learning approaches (Shen et al., 2018). Although these methods are effective and easily incorporated into DNNs, they are mainly designed for feature-based distribution shift alignment, where the primary goal is to align feature representations between domains. However, in many healthcare and biosignal applications, the dominant source of mismatch lies not only in the input features but in the label space itself. Consequently, models optimised purely for feature-space alignment struggle to generalise when the quality of labels differ. Therefore, unlike most discrepancy-based methods, our method primarily addresses label distribution shift problems where the source domain has large-scale, noisy labels.

4.1.2.2 Adversarial-based DA

Most adversarial-based methods are motivated by a theory which suggests that a good cross-domain representation contains no discriminative information about the origin of the input and, moreover, shows good performance to reduce domain discrepancy (Wang and Deng, 2018). Among these methods, the Domain-Adversarial Neural Network (DANN) (Ganin et al., 2016) first introduced adversarial training to domain adaptation. DANN utilises a shared feature extractor to learn feature embedding and a discriminator to maximise the domain difference. After the convergence of the whole training, general features are learnt for each input while their domain information cannot be discriminated. Following this, Domain Separation Networks (DSNs) (Bousmalis et al., 2016) and Adversarial Discriminative Domain Adaptation (ADDA) (Tzeng et al., 2017) propose more sophisticated feature extraction methods using shared and private feature extractors for task classifier and domain discriminator. Moreover, Dual Adversarial Domain Adaptation (DADA) (Du et al.,

2020) uses two discriminators stacked against each other to aid discriminative training.

The aforementioned methods mainly focus on discriminating domains as a coarse binary classification task while neglecting the rich information of the domain distribution. Moreover, most applications of DA focus mainly on medical image segmentation and classification tasks (Perone et al., 2019; Mei et al., 2020; Venkataramani et al., 2018). Our work aims to resolve the label distribution shift while keeping the input data constant for regression targets when analysing biosignal time series data. We propose a novel DA framework (UDAMA) with multi-discriminators to learn the coarse-grained and fine-grained domain information and proceed to validate it on a CRF prediction task.

4.2 Cardio-respiratory fitness prediction

CRF is one of the strongest predictors of CVD compared to other risk factors such as hypertension and type 2 diabetes (Laukkanen et al., 2001). CRF is routinely assessed via measurement of VO_2max, widely considered to be the benchmark measurement, which provides valuable insight into a person's overall fitness. However, obtaining gold-standard VO_2max measurements, is time-consuming and thus rarely performed in clinical settings, as it requires participants to undergo a maximal exercise test to exhaustion on a treadmill, while wearing a mask connected to a computerised gas analysis system to monitor ventilation and expired gas fractions.

Recently, less accurate measurement schemes, such as sub-maximal silver-standard exercise tests utilising modern wearables embedded with accelerometers and ECG sensors, have opened the door to monitoring population-level fitness. However, this alternative measurement method has been shown to demonstrate a measurement bias ranging from -3.0 to -1.6 ml $O_2/min/kg$ and a Pearson's r ranging from 0.57 to 0.79 (Gonzales et al., 2020b) relative to the gold standard. Aside from producing less accurate VO_2max values, these measurements also exhibit distribution mismatches, making it difficult to integrate them into clinical practice.

Similar to other healthcare applications, the distribution shift between silver and gold-standard labels in the CRF prediction task is often ill-defined. To tackle this issue, this thesis aims to adapt the source domain, characterised by noisy, yet large-scale silver-standard labels, to the target domain with small-scale gold-standard datasets. Specifically, we introduce a novel adversarial-based unsupervised domain adaptation framework with multiple domain discriminators (*i.e.*, UDAMA) to learn domain-invariant features and improve model validation via gold-standard VO_2max prediction.

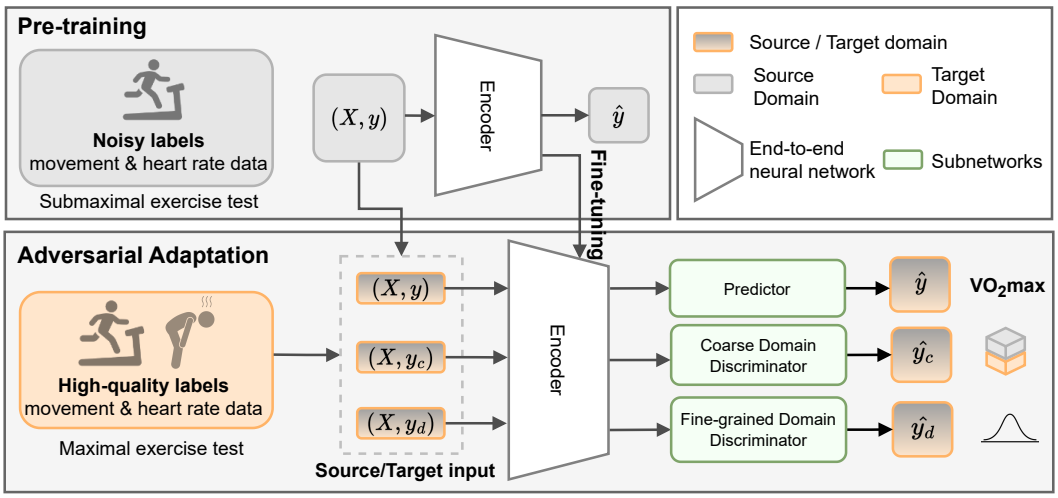


Figure 4.2: **UDAMA architecture and data pipeline:** We first pretrain the encoder on large-scale noisy silver-standard data. During adversarial adaptation, a small set of source samples is reintroduced to create a mixed-domain setting. The coarse-grained discriminator enforces global domain invariance, while the fine-grained discriminator aligns the mean and variance of the label distributions. Together, these components enable the encoder to learn representations that are both predictive on the target domain and robust to label distribution discrepancies.

4.3 Methods

This work introduces a novel unsupervised domain adaptation framework that utilises a multi-discriminator during adversarial training. The overall model architecture and multi-discriminator training scheme are shown in Figure 4.2. Herein, in this section, we formally discuss the problem formulation (Section 4.3.1) and details of our framework, which includes the first-step pretraining and second-step multi-discriminator domain adaptation training (Section 4.3.2).

4.3.1 Problem formulation and notation

Here, we denote $\boldsymbol{D_s}$ as the source domain containing silver-standard labels, and $\boldsymbol{D_t}$ as the target domain with gold-standard labels, as shown in Figure 4.2. For each domain, we assume the data in the form $\boldsymbol{X} = (\boldsymbol{x_1}, ..., \boldsymbol{x_n}) \in \mathbb{R}^{\boldsymbol{N} \times \boldsymbol{T} \times \boldsymbol{F}}$ corresponds to the accelerometer and ECG data from a chest-mounted device, using target regression VO_2max labels $\boldsymbol{y} = (\boldsymbol{y_1}, ..., \boldsymbol{y_n}) \in \mathbb{R}^{\boldsymbol{N}}$.

Additionally, we take into account contextual information, including height and weight, as metadata $\boldsymbol{M} = (\boldsymbol{m_1}, ..., \boldsymbol{m_n}) \in \mathbb{R}^{\boldsymbol{N} \times \boldsymbol{F}}$. For the input data $\boldsymbol{X}$ and $\boldsymbol{M}$, $\boldsymbol{N}$ represents the number of samples/subjects, $\boldsymbol{T}$ the length of input sequences, and $\boldsymbol{F}$ the number of input features. Besides, we use coarse- and fine-grained domain labels to train our model during adaptation and utilise multiple discriminators to differentiate between them. In

Table 4.1: **Notation.**

Notation	Description
$\boldsymbol{D_s}$	source domain with noisy (silver-standard) labels
$\boldsymbol{D_t}$	target domain with high-quality (gold-standard) labels
$\boldsymbol{X} \in \mathbb{R}^{\boldsymbol{N}\times\boldsymbol{T}\times\boldsymbol{F}}$	input time-series sequences
$\boldsymbol{M} \in \mathbb{R}^{\boldsymbol{N}\times\boldsymbol{F}}$	input user metadata
$\boldsymbol{N}$	number of samples
$\boldsymbol{T}$	length of input sequences
$\boldsymbol{F}$	number of features
$\boldsymbol{y_c} \in \mathbb{R}^{\boldsymbol{N}}$	categorical value representing the coarse-grained binary domain label
$\boldsymbol{y_d} \in \mathbb{R}^{\boldsymbol{N}}$	numerical value that denotes the fine-grained domain distribution label
$\boldsymbol{y} \in \mathbb{R}^{\boldsymbol{N}}$	numerical scalar for task target value
$\boldsymbol{D_c}$	coarse-grained domain discriminator
$\boldsymbol{D_f}$	fine-grained domain discriminator
$\boldsymbol{G_y}$	regression predictor
$\boldsymbol{E}$	feature encoder

particular, $\boldsymbol{y_c} = (\boldsymbol{y_c}[1], ..., (\boldsymbol{y_c}[n])$ is the categorical value representing the coarse-grained binary domain label. $\boldsymbol{y_d} = (\boldsymbol{y_d}[1], ..., (\boldsymbol{y_d}[n])$ is the numerical value that denotes the fine-grained domain distribution label. The coarse-grained domain discriminator is $\boldsymbol{D_c}$ and the fine-grained domain discriminator is $\boldsymbol{D_f}$. Also, for the training process, we denote the feature encoder with $\boldsymbol{E}$ and the regression predictor with $\boldsymbol{G_y}$. The overall networks thus can be represented as $\hat{\boldsymbol{y}}_{\boldsymbol{c}} = \boldsymbol{D_c} \cdot \boldsymbol{E}$, $\hat{\boldsymbol{y}}_{\boldsymbol{d}} = \boldsymbol{D_f} \cdot \boldsymbol{E}$ and $\hat{\boldsymbol{y}} = \boldsymbol{G_y} \cdot \boldsymbol{E}$. The full table is shown in Table 4.1.

4.3.2 Unsupervised domain adaptation and multi-discriminator adversarial training

Domain adaptation is a method for learning a mapping between domains with distinct distributions, including data distribution shifts such as covariate shift, conditional shift, and label distribution (Wang and Deng, 2018). In this chapter, we propose UDAMA, the unsupervised domain adversarial training, to address the label distribution shift problem, particularly when the source domain contains numerous noisy labels.

As shown in Figure 4.2, after pretraining on the source domain with large-scale silver-standard labels, we first incorporate part of prior knowledge from the $\boldsymbol{D_s}$ to create the adversarial training environment. Then we use the mixed silver-standard and gold-standard data to train the predictors and discriminator during the adaptation phase.

In particular, the adversarial training process consists of an encoder ($\boldsymbol{E}$), a VO_2max label predictor($\boldsymbol{G_y}$), and two domain classifiers/discriminators ($\boldsymbol{D}$) designed to solve label shift problems by distinguishing both the domain and domain distribution information.

During training, the fine-grained discriminator ($\boldsymbol{D_f}$) and coarse-grained ($\boldsymbol{D_c}$) discriminator are first optimised to identify the domain of each sample (*i.e.*, max$\boldsymbol{D_f, D_c}$). In adversarial training, the label predictor and encoder are then optimised to predict continuous fitness values from encoding(*i.e.*, min$\boldsymbol{E, G_y}$). The aforementioned adversarial process will finally achieve the trade-off (*i.e.*, the best prediction result in the most difficult-to-distinguish domain).

4.3.2.1 Coarse-grained discriminator

The coarse-grained discriminator ($\boldsymbol{D_c}$) is similar to other DAs (Mathelin et al., 2020; Zhao et al., 2018) and follows the DANN (Ganin et al., 2016) approach. In other words, $\boldsymbol{D_c}$ aims to discriminate the source of each data point in the mixture of pretraining and target labelled data as a binary classification task, where 0 represents the data comes from the $\boldsymbol{D_s}$, and 1 from $\boldsymbol{D_t}$. Specifically, after obtaining the representation matrix by fine-tuned feature extractor, two fully connected layers with the corresponding activation in $\boldsymbol{D_c}$ are used to discriminate the rough binary domains labels and predict a probability vector $\hat{\boldsymbol{y}}_{\boldsymbol{c}}$. Let $\hat{\boldsymbol{Y}}_{\boldsymbol{c}} = \hat{\boldsymbol{y}}_{\boldsymbol{c}}[\boldsymbol{n}]$ denote the predicted probability vectors for all the data points in ($\boldsymbol{D_t}$). The classification loss of the coarse-grained discriminator is then defined as:

$$L_{CSE} = \sum_{N} l_c(y_c[n], \hat{y}_c[n]) \tag{4.1}$$

where l_c is the cross entropy loss of a single data point, and by optimising $\boldsymbol{L_{CSE}}$ for $\boldsymbol{D_c}$, we can force the extractor to learn a general feature by maximising such divergence.

4.3.2.2 Fine-grained discriminator

However, a simple binary classification task cannot fully and accurately represent the domain label distribution. Therefore, we augment adversarial training with a fine-grained discriminator ($\boldsymbol{D_f}$) to discriminate distribution differences. Specifically, instead of generating the binary domain labels (0 or 1) for the source or target domain for each sample, we construct a more complex pseudo-label ($\boldsymbol{y_d}$) to represent its domain label distribution. Based on our observation of the health outcome labels, which conform to a Gaussian distribution (as shown in Figure 4.1), we then assign the ($\boldsymbol{y_d}$) for each training sample during adaptation. Specifically, $\boldsymbol{y_d}$ represents the mean and variance of the regression label distribution. This label is based on whether the sample is from the pretraining or target domain.

Therefore, after generating the feature matrix using $\boldsymbol{E}$, $\boldsymbol{D_f}$ is designed to distinguish the mean and variance of the label distribution using two fully connected layers with the corresponding activation. Let $\hat{\boldsymbol{Y}}_{\boldsymbol{d}} = \hat{\boldsymbol{y}}_{\boldsymbol{d}}[\boldsymbol{n}]$ denote the predicted probability vectors for all

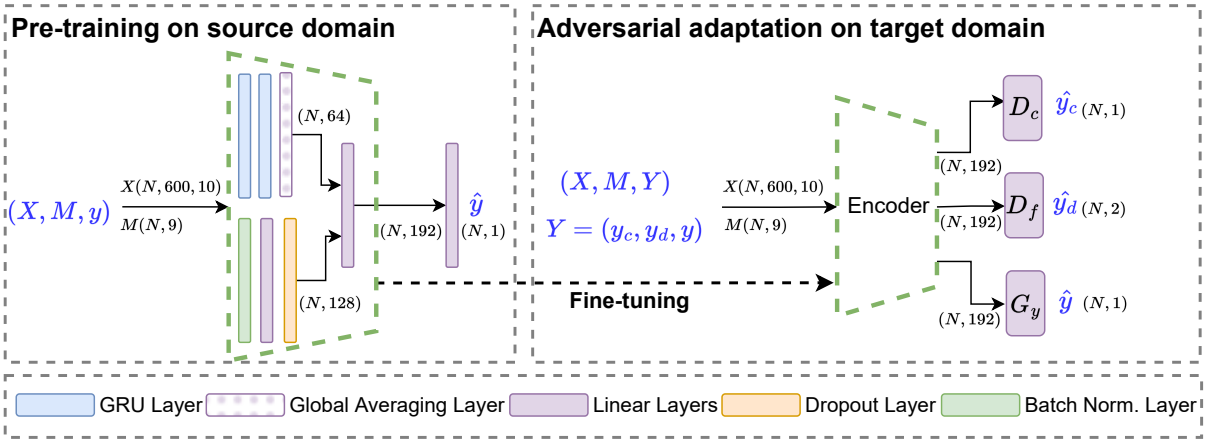


Figure 4.3: **UDAMA detailed model architecture.** Visualisation of the detailed network structure and the input and output dimensionality for each module.

the data points in $(\boldsymbol{D_t})$. Then, the loss of the fine-grained discriminator is defined as:

$$L_{GLL} = \sum_{N} l_g(y_d[n], \hat{y}_d[n]) \tag{4.2}$$

where l_g is each data point's Gaussian Negative Log-Likelihood (GLL) loss. In particular, GLL optimises the mean and variance of a distribution and thus further maximises the nuance changes among the sample and updates the discriminator.

After this process, D_c and D_f can maximise the difference arising between the source and target domains for the multiple-domain discriminator training scheme. Meanwhile, the encoder and the predictor try to maximise the correct prediction of y. These two modules play two games and finally reach a balance during the training. As a result, the encoder and predictor can learn a representation that cannot tell the difference between the source and target domains after the training has converged.

4.3.2.3 Objective functions and training

For adversarial training, the shared encoder first leverages the pretrained model and then extract general features. Following this step, discriminators are trained simultaneously to differentiate the domain labels. The predictor is used for the regression healthcare outcomes prediction task, and a mean squared error loss $\boldsymbol{L_{MSE}}$ is applied to optimise the $\boldsymbol{G_y}$. The detailed training overflow and input for each module are shown in Figure 4.3.

Finally, the whole framework can be optimised by the total loss $\boldsymbol{L}$, which is defined as:

$$L = \alpha L_{MSE} - \lambda_1 L_{CSE} - \lambda_2 L_{GLL} \tag{4.3}$$

We optimise the overall loss $\boldsymbol{L}$ so as to minimise the predictor loss while maximising the loss of domain discriminators. In detail, α is used to scale down the predictor loss to the

same level as predictors, λ_1 and λ_2 control the relative weight of the discriminator loss, where $\lambda_1 + \lambda_2 = 1$. Details of hyperparameter choice are discussed further in Section 4.4.

4.4 Experiment setup

We conduct various qualitative and quantitative evaluations to validate our models on the CRF prediction task with wearable sensing (Section 4.4.1). We outline the model's architecture and how it differs from other baselines (Section 4.4.3 and Section 4.4.4). Next, we further quantify the impact of incorporating source domain knowledge to UDAMA (Section 4.4.5). The overall workflow is depicted in Figure 4.2.

4.4.1 Datasets and training strategy

As discussed in Section 2.2, we employed the Fenland and BBVS datasets for the cardio-fitness prediction task. The details of these datasets and the corresponding evaluation settings are provided in this section. The target domain $\boldsymbol{D_t}$ represents the gold-standard measurement dataset BBVS, a subset of 181 participants from the Fenland study with directly measured gold-standard VO_2max (Gonzales et al., 2021) from maximal exercise tests. In the BBVS study, participants were required to wear a face mask to quantify respiratory gas measurements (Rietjens et al., 2001) to exhaustion.

Fenland Study. The Fenland dataset with large-scale weakly-labelled VO_2max is used as the source domain (D_s) in this chapter. In particular, the Fenland study (Lindsay et al., 2019b) is a prospective population-based cohort study of individuals aged 35-65 which investigated the interaction between environmental and genetic factors in determining predispositions to obesity, type 2 diabetes, and related metabolic disorders. The study collected data from 12,435 participants in Cambridgeshire between 2005 and 2015. After a baseline visit to the clinic, participants were instructed to conduct a validated sub-maximal treadmill test to generate silver-standard labelling data in relation to their VO_2max using a linear regression method while wearing the chest ECG sensor Actiheart for six consecutive days to concurrently collect their heart rate and movement data. Data from 11,059 participants were included in this study following the exclusion of those with insufficient or corrupt data and/or missing variables.

Biobank Validation Study (BBVS). The BBVS dataset serves as the target domain (D_t) with clean gold-standard VO_2max labels (Gonzales et al., 2021). The BBVS dataset is a subset of 191 participants from the Fenland study, from which gold-standard VO_2max data were obtained. Following standardised techniques used for the UKB-CRF test (UK Biobank, 2011), participants performed 5 maximal exercise tests to elicit VO_2max for six

consecutive days, as per the Fenland study. In the BBVS study, participants were required to wear face masks with a computerised metabolic system to measure their VO_2max during the study. ECG data were obtained via (Cardiosoft) using an Actiwave CARDIO device (CamNtech, Papworth, UK). Similarly, data from 181 participants were included after excluding those participants with insufficient data. All participants from the two datasets provided written informed consent, and the study was approved by the University of Cambridge Ethics Committee.

4.4.2 Data preprocessing and feature extraction

For both the Fenland and BBVS studies, participants wore the two standard ECG Actiheart devices attached to the chest, which recorded both heart rate and movement at 60-second intervals (Brage et al., 2005) via wearable free-living sensing. All heart rate (HR) data collected during laboratory visits underwent pre-processing for noise filtering (Stegle et al., 2008). Participants were excluded if they had fewer than 72 hours of concurrent wear data (*i.e.*, three full days of recording) or lacked adequate treadmill-based calibration data. Non-wear periods were detected and removed using standard procedures, where extended intervals of non-physiological HR or no movement (>90 minutes, as reported by the device accelerometer) were flagged as non-wear.

To standardise movement intensity, we converted accelerometer outputs into metabolic equivalents (METs) using the formula:

$$1 \text{ MET} = 71 \text{ J}\,\text{min}^{-1}\,\text{kg}^{-1} = 3.5 \text{ ml}\,\text{O}_2\,\text{min}^{-1}\,\text{kg}^{-1}$$

Behaviours were then classified into activity levels, as follows: sedentary (<1.5 METs), moderate-to-vigorous physical activity (MVPA; 3–6 METs), and vigorous physical activity (VPA; >6 METs). To account for daily rhythms, sensor timestamps were encoded using cyclical temporal features (Spathis et al., 2021).

Given the high sampling rate (1 sample/min), week-long sequences typically contained over 10,000 time steps, making direct modelling computationally infeasible. To reduce sequence length, we downsampled both HR and accelerometer signals by a factor of 15, resulting in sequences of approximately 600 time steps. All features were subsequently normalised using min–max scaling (sequence-wise for sensor data and column-wise for metadata).

In total, each input feature vector comprised 19 features combining both time series and metadata, which were then used as inputs to various DNNs. A detailed description of these features is provided in Table 4.2.

Table 4.2: **Description of the features/variables used in our analysis as inputs to the models.** The features with asterisks(*) are time-series and remaining are metadata. The final set of features is 19.

Features/Variables	Description
Sensors	
Acceleration*	Acceleration measured in mg
Heart rate (HR)*	Mean HR resampled in 15sec intervals, measured in BPM
Heart Rate Variability (HRV)*	HRV calculated by differencing the second-shortest and the second-longest inter-beat interval (as seen in Faurholt-Jepsen et al. (2017)), measured in ms
Acceleration-derived Euclidean Norm Minus One (ENMO)*	ENMO-like variable (Acceleration/0.0060321) + 0.057 (as seen in White et al. (2016))
Acceleration-derived Metabolic Equivalents of Task (METs)*	
Sedentary*	If Accelerometer <1, take daily count and average
Moderate to Vigorous*	If Accelerometer >= 1, take daily count and average
Vigorous*	If Accelerometer >= 4.15, take daily count and average
Anthropometrics	
Age	Age, measured in years
Sex	Sex is binary (female/male)
Weight	Weight, measured in kilograms
Height	Height, measured in meters.centimeters
Body Mass Index (BMI)	BMI is calculated by Weight/(Height^2), measured in kg/m^2
Resting Heart Rate	
Wearable-derived RHR	RHR is calculated by averaging the 4th, 5th, and 6th minute of the baseline visit and adding to that the Sleeping Heart Rate that has been inferred by the wearable device. Gonzales et al. (2020a)
Seasonality	
Month of year	The month number is used along with a coordinate encoding that allows the models to make sense of their cyclical sequence.

4.4.3 Model architecture and tuning

Training strategy. We evaluate UDAMA on these two datasets by first pretraining a model on the $\boldsymbol{D_s}$ Fenland. Second, we develop the adversarial training framework with multi-discriminators on the BBVS ($\boldsymbol{D_t}$), with the help of incorporated prior domain knowledge (*i.e.*, injecting random samples from the source domain). After the adversarial adaptation, we predict VO_2max on the held-out test set of BBVS using 3-fold cross-validation using UDAMA. Within each fold, the dataset is split into 70% training and 30% testing consisting only of target domain samples.

Model architecture. This section discusses the details of the NNs used in this study. For the encoder network, two modules are integrated within the network to extract temporal and metadata information. In particular, we use two Bidirectional GRU layers of 32 units for the time series data module, followed by one 1D-global averaging pooling layer. On the other hand, in the metadata module, MLP layers of dimensionality 128 are constructed to extract associated metadata representation after a Batch Normalisation layer. Following this, the time and metadata module outputs were concatenated together to generate a complete embedding matrix for the subsequent regression or classification tasks. The training architecture employed was as shown in Figure 4.3.

The training pipeline consists of two phases using the pretraining and fine-tuning learning scheme. First, after comparing different parameter-sharing techniques for fine-tuning the encoder utilised in the source domain, we froze the first GRU and MLP layers and fine-tuned the remaining network. When compared with freezing all layers except re-training the output layer, the fine-tuning scheme in our network could capture the general features from the lower layers and extract problem-specific characteristics from higher layers. Lastly, the representation embedding produced from the fine-tuned encoder is transmitted to distinct tasks with a linear activation layer appropriate for predicting the fitness level or classifying the domain labels.

Hyper-parameter tuning. All network blocks in the framework are trained using Adam with a learning rate tuned over $\{1e^{-2}, 1e^{-3}\}$. The dropout rate is tuned over the following ranges {0.2, 0.3}. Moreover, we tuned the batch size between 8, 16, and 32, based on the efficiency and stability of the training process. To tune the UDAMA total loss, we conducted a grid search {0.01, 0.02, 0.03} for the α and {(0.9, 0.1), (0.8, 0.2), (0.7, 0.3), (0.6, 0.4), (0.5, 0.5)} for the combination of λ_1 and λ_2. We performed early stopping to combat overfitting until the validation loss stopped improving after ten epochs. The details of hyperparameter selection are as listed in Table 4.3.

Table 4.3: **Hyper-parameter tuning**.

Parameter	**Search space**	**Selected value**
Dropout	{0.2,0.3}	0.3
optimiser	Adam	Adam
Learning rate	{1e-2, 1e-3}	1e-3
Epochs	[0,100]	Early stopping
Batch size	{8,26,32}	8
α	{0.01,0.02,0.03}	0.01
λ	{(0.9, 0.1), (0.8, 0.2), (0.7, 0.3), (0.6, 0.4), (0.5, 0.5)}	(0.9,0.1)

4.4.4 Baselines

To verify the effectiveness of our proposed network, we compare the UDAMA against recent SOTA baselines (as introduced in Section 2.4 and Section 4.1):

- **In-domain supervised model.** A multi-model network with training on the same domain train and test set.
- **Out-of-domain supervised model**. A network with the same structure as the pretraining model, using wearable data and common biomarkers in $\boldsymbol{D_t}$ as inputs to predict VO_2max.
- **Transfer learning.** A pretrained model trained on $\boldsymbol{D_s}$ is reused and fine-tuned on the target domain.
- **Autoencoder (Srivastava et al., 2015).** pretrain a model with stacked recurrent autoencoders on $\boldsymbol{D_s}$ and fine-tune the representation from the encoder to $\boldsymbol{D_t}$.
- **Deep-Coral (?)**. A widely employed discrepancy-based domain adaptation minimises the divergence between the source and target in feature space. In particular, it seeks to align the second-order statistics of the source and target distributions.
- **WDGRL (Shen et al., 2018)**. Wasserstein Distance is used to minimise the disparity between the source and target representations in the feature space. Specifically, the distance will be optimised in an adversarial way for the feature extractor, and domain-invariant features will be learned.
- **Domain Adversarial Neural Networks (DANN) (Ganin et al., 2016).** A benchmark domain adaptation method that uses adversarial training for binary domain classification.

Although most recent domain adaptation methods have shown enhanced performance in dealing with distribution shift through self-training or sophisticated adversarial training schemes (Du et al., 2020; Liu et al., 2021), they do not specifically tackle the regression

tasks or healthcare datasets that contain both noisy and gold-standard labels. Therefore, their relevance in the cardio-fitness prediction task is limited.

4.4.5 Effect of injected source domain samples

In adversarial-based domain adaptation, the training environment is constructed using both source and target domain samples, whereupon the domain discriminator is exposed to inputs from both domains, while the feature extractor is optimised to produce domain-invariant representations. Our proposed approach leverages prior knowledge from the source domain to strengthen this adversarial environment.

To assess the degree to which the incorporated source domain knowledge impacts our model, we evaluate UDAMA on $\boldsymbol{D_t}$ with different levels of injected samples from ($\boldsymbol{D_s}$). As such, we put {0.1%, 0.2%, 0.4%, 1%, 2%, 4% } from $\boldsymbol{D_s}$, which equals {1%, 5%, 10%, 30%, 50% and 100%} of the training data of the target domain ($\boldsymbol{D_t}$). We set the maximum amount to 100% because the two domains are highly divergent and, for any larger amount, the model would be trying to predict the silver data rather than the gold data. Specifically, the number of samples differs between source and target domain samples (source/target = 25:1), which means that 0.2% of the source data is equivalent to 5% of the target data. If we continually add more source data, for example, 4% of the source (which equals 100% of the target data), we will ultimately train on the source data instead of the target data. Finally, we ran the experiments 15 times with different seeds to assess the impact of the injected noisy samples using the average performance on the VO_2max prediction task. The results are shown in Section 4.5.3.

4.4.6 Metrics

The prediction performance of all models is evaluated based on standard regression evaluation metrics such as the mean squared error (MSE) and mean absolute error (MAE). The coefficient of determination (R^2) and the Pearson correlation coefficient (Corr) are also used to evaluate the model's performance on the health-related outcome prediction task.

4.5 Results

In this section, we present the results of applying UDAMA to the cardio-fitness prediction task and compare these to the baselines (Section 4.5.1). Additionally, we discuss the performance in addressing the domain shift problem (Section 4.5.2) and the impact of injected source domain knowledge (Section 4.5.3). Furthermore, we conduct an ablation

Table 4.4: **Evaluation of different methods on the CRF prediction task.** Each result displays the mean value with standard deviation from three-fold cross-validation. In particular, **In-domain** means the model is trained on the same domain, namely trained on D_t and tested on D_t, while **Out-of-domain** corresponds to models trained on Fenland (D_s) and adapted/fine-tuned to BBVS (D_t). All **Out-of-domain** models are evaluated on the BBVS test set. UDAMA substantially achieves the best performance across all four metrics over all out-of-domain baselines and even surpasses in-domain supervised training, demonstrating its strong ability to adapt from noisy source data to the gold-standard target domain.

In-domain	**Training method**	$\mathbf{R^2}$	**Corr**	**MSE**	**MAE**
$D_t \rightarrow D_t$	*Supervised*	0.123 ± 0.111	0.622 ± 0.036	43.778 ± 6.012	5.263 ± 0.277
Out-of-domain	**Training method**	$\mathbf{R^2}$	**Corr**	**MSE**	**MAE**
$D_s \rightarrow D_t$	*Supervised*	-0.096 ± 0.100	0.007 ± 0.250	58.048 ± 10.061	6.336 ± 0.621
	WDGRL (Shen et al., 2018)	-0.100 ± 0.073	0.004 ± 0.161	55.611 ± 10.61	6.044 ± 0.615
	Autoencoder (Srivastava et al., 2015)	-0.067 ± 0.069	0.127 ± 0.222	53.254 ± 4.878	5.973 ± 0.194
	Deep-Coral (Sun and Saenko, 2016)	0.021 ± 0.073	0.360 ± 0.057	49.044 ± 6.553	5.638 ± 0.374
	Transfer learning (TF)	0.283 ± 0.037	0.621 ± 0.012	35.399 ± 5.910	4.744 ± 0.433
	DANN (Ganin et al., 2016)	0.288 ± 0.077	0.617 ± 0.037	35.458 ± 3.920	4.679 ± 0.382
	UDAMA (ours)	$\mathbf{0.459 \pm 0.063}$	$\mathbf{0.701 \pm 0.032}$	$\mathbf{27.469 \pm 6.456}$	$\mathbf{4.111 \pm 0.353}$

study to verify the effectiveness of our framework's structure (Section 4.5.4). Finally, we discuss the model robustness in Section 4.5.5 with semi-synthetic data.

4.5.1 Fitness prediction

We took 60 participants from the BBVS dataset as test samples to predict their CRF by predicting VO_2max values. The comparison between the proposed domain adaptation framework and baseline approaches is shown in Table 4.4.

First, in the Out-of-domain comparison, adversarial-based DA or transfer learning shows better performance than the discrepancy-based method under label distribution shift. In particular, the discrepancy-based method, Deep-coral, increases the Corr and MAE to 0.36 and 5.638, respectively. Meanwhile, WDGRL aims to minimise the feature difference by employing the Wasserstein distance, yielding results comparable to the Out-of-domain supervised method, which directly applies the model trained on Fenland for testing BBVS. In contrast, the adversarial-based methods employed here display better results. DANN learns a representation that is predictive of the regression task but uninformative to the input domain and improves the Corr and MAE to 0.617 and 4.679, respectively. According to Corr and MAE, methods such as transfer learning with fine-tuning techniques also improve performance, achieving 0.621 and 4.744, respectively, relative to the discrepancy-based method.

In general, high Corr and R^2 values and low MSE and MAE demonstrate the model's ability to leverage noisy, large-scale labelled VO_2max data for gold-standard VO_2max prediction under label shift. However, the limited size of the test set might result in

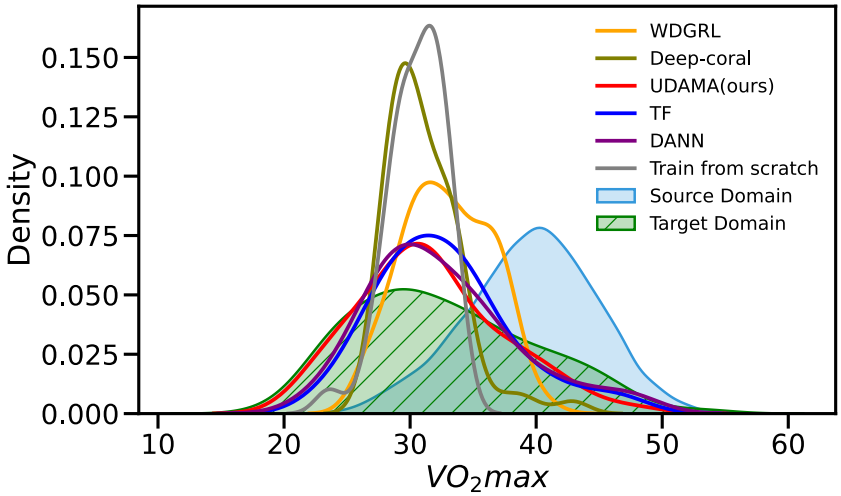


Figure 4.4: **Predicted VO_2max distributions of the BBVS test set:** Most baseline methods fail to recover the BBVS distribution, highlighting the challenge of adapting from noisy, mismatched source data

increased uncertainty during model evaluation. The results from both TF and DANN are similar, as shown in Table 4.4. In contrast, UDAMA utilising both fine-tuning and adversarial-based domain adaptation methods outperforms all the aforementioned baselines. We observe that the correlation (Corr) outperforms the basic transfer learning methods by 12.9%, the MSE increases the gain in performance by 22.4%, and R^2 facilitates an improvement of 62.2%. Moreover, compared with the in-domain supervised training on $\boldsymbol{D_t}$, UDAMA shows a significant increase, improving R^2 from 0.123 to 0.459. Our method also achieves good performance when generalising the model from the in-domain to out-of-domain BBVS setting, compared with the dramatic drop in performance, as shown in Table 4.4. Therefore, UDAMA can leverage information from large-scale, noisy datasets and alleviate model performance degeneration relative to directly validating models on small-scale sensing datasets.

4.5.2 Domain shift

To better establish whether UDAMA or baseline DA methods can solve the label distribution shift problem effectively, Figure 4.4 presents the predicted label distribution of the BBVS test set obtained using different methods. First, it shows that the $\boldsymbol{D_s}$ dataset (*i.e.*, Fenland) shares a different underlying VO_2max distribution when compared to $\boldsymbol{D_t}$ BBVS. Our results demonstrate that UDAMA can learn the small dataset distribution during the adaptation phase and achieve promising results relative to other methods. Further, we observe that both adversarial-based methods and transfer learning capture the mean and range of the target domain distribution, as shown in Figure 4.5, whereas discrepancy-based methods fail to learn the general distribution. We attribute this performance degeneration

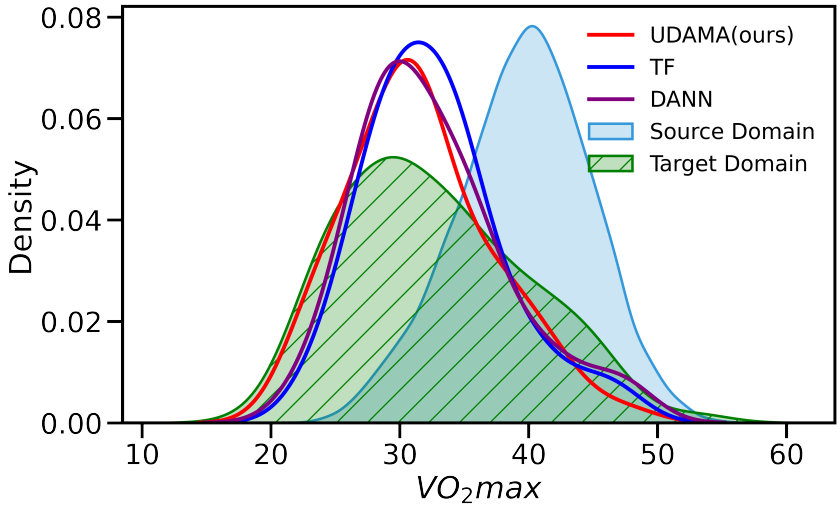


Figure 4.5: **UDAMA mitigates distribution shifts.** UDAMA closely matches the target BBVS VO_2max distribution compared with TF and DANN.

to the fact that discrepancy-based approaches, which are largely designed to minimise the divergence arising between feature spaces, cannot alleviate the impact of noisy labelling.

We also noted that, as shown in Figure 4.5 and Figure 4.4, DANN and TF yield broadly similar distributions. However, these plots mainly show that both methods capture the global mean and range of the target distribution, without sufficiently aligning the fine-grained structure. Therefore, we use the Hellinger Distance (HD), which calculates the similarity of distributions between prediction and ground truth to examine the distance between two label distributions and provides more fine-grained local density information. Specifically, our framework's prediction of fitness level lies within the same range as the ground truth, while methods like Deep-coral or WDGRL fail at learning within this range. Besides, the distribution of UDAMA ties is close when compared to baseline methods, where the normalised HD for UDAMA is 0.179, the HD for TF is 0.264, and for Deep-coral is 0.305. These results indicate that our framework effectively alleviates the distribution shift problem inherent in the VO_2max prediction task and UDAMA can leverage noisy silver-standard data to improve the performance on the gold-standard dataset.

4.5.3 Impact of injected knowledge from source domain

Instead of using a generator, we incorporated source domain knowledge to create an adversarial training environment. Figure 4.6 shows the average performance of adding the different scales of injected source domain samples to the target domain. Each box plot shows the average MSE results obtained from 15 runs with added random samples from the source domain. As shown in Figure 4.6, our method performs better than the baseline methods with added samples in all cases, even after adding 100% of the noisy samples of $\boldsymbol{D_t}$ from $\boldsymbol{D_s}$. Specifically, we observe that it achieves the best results and showcases the

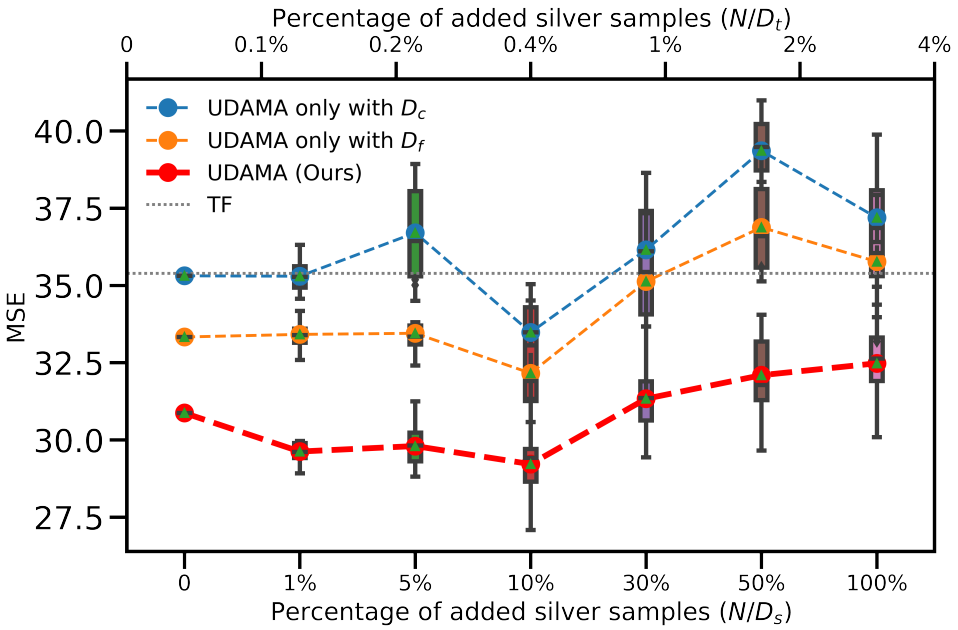


Figure 4.6: **Impact of injected source domain samples.** UDAMA achieves the strongest positive transfer when only a very small portion of source samples is injected, while larger injections gradually degrade performance as noisy source data begins to dominate the adaptation process.

most positive transfer by only adding 0.4% of the source domain, equivalent to 10% of $\boldsymbol{D_t}$. In contrast, if we continue to add more noisy information to the adaptation stage, the performance will gradually decrease, as the source and target domain data reach a ratio of 1:1. After that, the adaptation tends to learn the noisy source domain representation rather than the target domain, yielding a negative transfer. As a result, injecting only a few samples from the source domain to create the adversarial training environment might help to learn more domain-invariant features and achieve optimal results. However, A natural question is whether the injected noisy–source ratio can be chosen in a more principled manner than empirical search. In practice, the optimal ratio depends on how fast the discriminator begins to overfit to the source distribution, which varies with the degree of domain divergence. Our stability analysis shows that very small injections (0.2–0.4%) consistently provide the strongest transfer signal without overwhelming the target domain. A more principled future direction would be to adaptively adjust the injection ratio during training using divergence- or uncertainty-based criteria, thereby reducing reliance on manual tuning.

4.5.4 Ablation study

Our framework comprises two joint discriminators, so we perform an ablation study to understand the effect of each discriminator. Based on the observation that incorporating

Table 4.5: **Ablation study by removing one of the discriminators. Both coarse- and fine-grained discriminators contribute positively, but the fine-grained discriminator alone can yield strong gains over baseline methods**

Discriminator	R^2	Corr	MSE	MAE
Coarse-grained	0.339 ± 0.053	0.626 ± 0.032	33.485 ± 6.740	4.580 ± 0.337
Fine-grained	0.394 ± 0.030	0.666 ± 0.030	30.578 ± 5.300	4.498 ± 0.324
UDAMA (ours)	**0.459 ± 0.063**	**0.701 ± 0.032**	**27.469 ± 6.456**	**4.111 ± 0.353**

a 10% amount of BBVS from the $\boldsymbol{D_t}$ Fenland dataset achieves the best transfer, we trained different discriminators under this setting. We observed that a single discriminator (coarse-grained or fine-grained) exhibits more competitive performance than the baseline TF or DANN methods, as shown in Table 4.5 and Figure fig:scale. Specifically, the fine-grained discriminator alone significantly outperforms the baseline results: +15.8% TF (based on MSE). This indicates that utilising a fine-grained distribution domain discriminator, which effectively discriminates the domain label distribution, enhances the training framework. Therefore, using a combination of $\boldsymbol{D_f}$ and $\boldsymbol{D_c}$, UDAMA can effectively learn cross-domain information without capturing the domain information of the input and thus alleviate the noisy labelling problem. A common limitation of adversarial training is its sensitivity to imbalance between discriminator and encoder updates, which can lead to unstable optimisation. In our setting, stability is supported by early stopping, grid-searched discriminator loss parameter and evaluation over 15 random seeds, all of which consistently yield low-variance results. Nonetheless, the framework still depends on careful tuning of discriminator weights, and future work could investigate adaptive or learnable weighting schemes to further mitigate sensitivity.

Additionally, our method is unique in that it uses a wearable-based CRF prediction task, unlike other methods that solely rely on anthropometric data (Nes et al., 2011). Our experiments have demonstrated that combining sensor data with anthropometric information leads to improved prediction accuracy. Although the sensor data alone is unreliable and insufficient for accurate VO_2max prediction, when combined with anthropometric data, the accuracy of UDAMA increases from 0.679 (using only anthropometric data) to 0.701 (using anthropometric data and wearable sensor data such as acceleration, heart rate, and heart rate variability). As a result, using anthropometric and low-cost wearable devices in conjunction enables more accurate VO_2max prediction through UDAMA.

4.5.5 Robustness assessment with semi-synthetic data shift

Although the distribution shift observed among gold and silver-standard labels is prevalent in healthcare applications, there are very few available open datasets to which we can apply our method. Therefore, to further evaluate the efficacy of UDAMA, we generated semi-

Table 4.6: **Simulated Label Distribution Shift**. Kullback–Leibler divergence (KL divergence) calculates the distribution difference between the shifted D_s and fixed D_t. UDAMA remains substantially more robust than DANN across all shift severities

Source dataset	**KL divergence**	**UDAMA Corr**	**DANN Corr**
Fenland	0.461	0.701 ± 0.032	0.617 ± 0.037
Left shifted Fenland	1.607	0.656 ± 0.065	0.589 ± 0.027
Right shifted Fenland	3.188	0.646 ± 0.035	0.608 ± 0.037

synthetic datasets with various label distribution shifts for the cardio-fitness prediction. Motivated by the label distribution shift simulation for classification (Lipton et al., 2018), we shift the labels in the $\boldsymbol{D_s}$ by a fixed offset and Gaussian noise. By conducting experiments with varying degrees of label shifts, we gain a comprehensive understanding of the effect of label shifts on CRF prediction tasks. As the shift becomes greater from the target domain, the performance decreases. In particular, we show two extreme cases by pushing the source domain shift to the left and right to stress-test UDAMA, and the results are shown in Table 4.6. Despite the increasing KL divergence, which indicates a greater deviation from the ground truth data, our method still displays robust performance relative to the baseline DA method, especially in the case of low fitness on the left side. These stress tests highlight the versatility and robustness of our model in dealing with different input distributions.

4.6 Discussion and conclusions

This chapter examined the challenge of building generalisable models for out-of-domain real-world healthcare applications, focussing on the problem of *label distribution shift*. Unlike the assumptions in the previous chapter and much of the existing literature, validated healthcare datasets rarely come with large-scale gold-standard annotations. In fact, gold-standard labels are expensive, difficult to obtain, and typically result in small and sparse datasets that limit generalisation to new cohorts and applications.

Modern wearable devices, on the contrary, can provide large-scale *silver-standard* labels at low cost. Although such labels offer scalability, they are inherently noisy and often exhibit distributional discrepancies compared to gold-standard ground truth. These discrepancies make it difficult to directly deploy models trained solely on silver-standard data. To bridge this gap, we introduced a multi-discriminator domain adaptation method for cross-domain representation learning, which transfers knowledge from large-scale weakly labelled data to small-scale health datasets with gold-standard labels. Specifically, in the context of CRF prediction, our results demonstrate that leveraging large-scale noisy VO2max labels using UDAMA not only achieves improved fitness prediction but also effectively mitigates label

distribution shifts. This paves the way for the practical application of machine learning for real-world health outcomes. Furthermore, our approach can be easily adapted to various health-related tasks, particularly those involving high-dimensional time-series data and changes in label distribution for regression tasks.

This second empirical study alleviates the issue of out-of-domain distribution shift by addressing challenges at the level of *annotations*. However, it still presents several limitations. First, adversarial training requires manually generating or injecting noisy sampled data and relies on sensitive hyper-parameter tuning to determine an effective noise ratio, underscoring the need for more automated or learnable adaptation strategies. Second, the empiracle study assess only under cardiovascular-related tasks and assumes the model have access to the label distribution. Second, the empirical study focuses exclusively on cardiovascular-related tasks and assumes access to the target label distribution. While the existing training framework shows effectiveness under this assumption, it may fail in unseen domains where the label distribution diverges from Gaussian assumptions. Furthermore, both the first and second contributions primarily focus on the same dataset or set of tasks.

Although these contributions represent important progress, they diverge slightly from the broader objective of this thesis: developing effective and generalisable models for biosignal time series. Beyond label-related challenges, biosignal modelling must also address data complexities, such as irregular sampling, missingness, and general multi-domain adaptation to unseen tasks and domains across heterogeneous datasets. The next chapter moves beyond label-related challenges and focusses on both data-centric and domain heterogeneity issues. Specifically, we present a framework for modelling irregular biosignal time series and their derived vital signs across multiple health-related tasks, enabling us to advance toward a universal representation learning paradigm for real-world biosignal data.

Chapter 5

FORM: A Foundation Model for Irregular Multi-domain Biosignal Time Series

In previous chapters, we proposed self-supervised frameworks to learn effective representations from large-scale unlabelled biosignal data, demonstrating strong performance and efficiency in diverse healthcare applications. We also introduced fine-grained domain adaptation methods to mitigate label distribution shifts when deploying models in real-world settings, ensuring that representations remain stable across heterogeneous cohorts and improving robustness when transferring models to new domains.

Despite the advances presented in the previous chapters, these methods still struggle to build effective and generalisable biosignal models. The challenge arises from the need to train in multiple heterogeneous cross-domain data sets and from inherent data characteristics, particularly missing values and irregular sampling, as discussed in Section 1.2. Specifically, real-world biosignal data often have non-uniform time intervals between successive observations and introduce missingness in the input data. For example, in clinical settings such as intensive care units (ICUs), patient physiological biosignal data, including heart rate, blood pressure, and respiratory rate, are frequently recorded at irregular intervals using various physiological sensors, where these sensors are triggered at irregular intervals by events in real life for each sensor. Accurate modelling of such irregularly sampled data, for example, using historical data from the past 24 hours to forecast patient conditions in the subsequent 24 hours, is essential for timely patient monitoring, clinical decision-making, and health status inference, as shown in Figure 5.1. However, most existing methods including SOTA methods for time series modelling in Chapter 3 & 4 assume regular biosignal time series with constant time intervals between all consecutive observations with all features observed at any given time (Marlin et al., 2012) and do not encode irregular time information during modelling. Therefore, simply

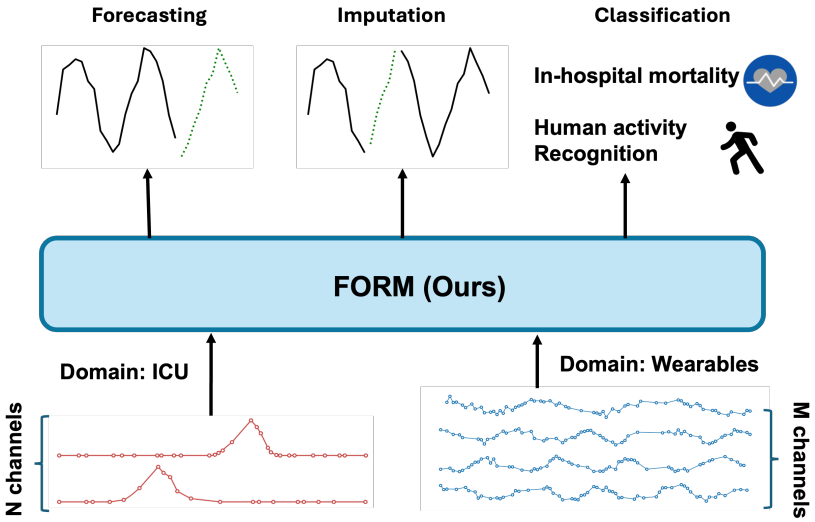


Figure 5.1: FORM learns representations from multiple datasets with varying numbers of input channels (variables) and irregularities and then generalises to different downstream tasks.

adapting regular time series methods for irregular biosignal time series is suboptimal, as these approaches fail to capture the uneven temporal structure and missingness patterns that fundamentally shape the underlying physiological dynamics.

To enable effective modelling for irregular time series, recent approaches fall primarily into four categories. *Attention-based models* (Shukla and Marlin, 2021; Zhang et al., 2019; Chen et al., 2023c) extend Transformer (Vaswani et al., 2017) architectures to capture irregularities in temporal dynamics. *ODE-based models* (Chen et al., 2018) use differential equations to model irregular and continuous-time dynamics. *Graph-based models* (Zhang et al., 2022c, 2024a) exploit the relational structure among multivariate irregular time series data, addressing complex inter-correlations, and *self-supervised models* (Chowdhury et al., 2023; Beebe-Wang et al., 2023) leverage intrinsic similarities within unlabelled irregular data, overcoming limitations posed by architectures traditionally reliant on fixed and regular sampling intervals (Yue et al., 2021; Wu et al., 2021).

Although effective in modelling irregular information, they are typically trained in a dataset-specific manner on the same datasets or domains, significantly limiting their generalisation capabilities to new domains with differing variable sets and missingness patterns. For example, models trained on 2-lead ECG data for heart rhythm classification often fail to generalise beyond that specific dataset, for example, to 12-lead ECG data. More importantly, such dataset-specific approaches hinder performance across various tasks and scenarios with scarce labelled training data.

To address the limitations mentioned above, large pre-trained models are needed to be trained on extensive, unlabelled, and irregular data from multiple biosignal datasets. Such

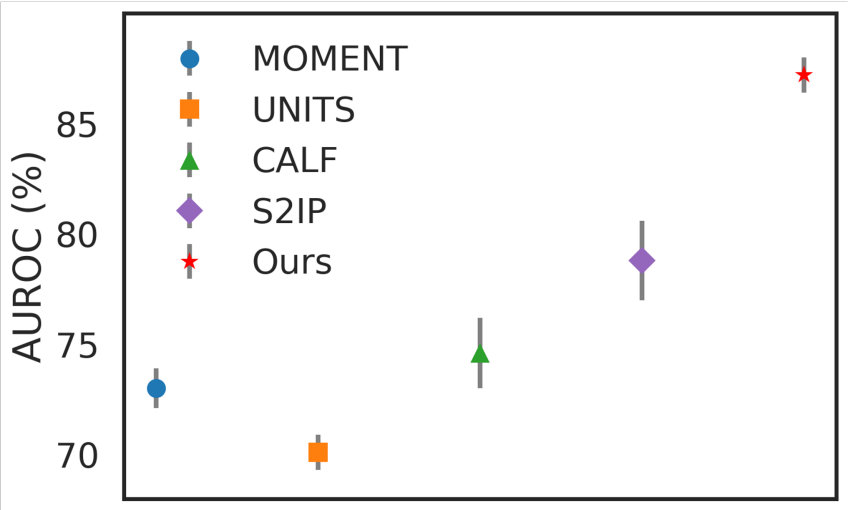


Figure 5.2: Existing foundation models for regular time series struggle to capture complex patterns inherent to irregular time series on the irregular Physionet dataset for the in-hospital mortality classification task.

models can then be adapted to downstream tasks through transfer learning and supervised fine-tuning, reducing reliance on scarce labelled data. In the time-series domain, large pretrained models typically follow a pretrain–fine-tune paradigm: they are first trained with self-supervised objectives (*e.g.*, masked reconstruction, contrastive learning, next-step prediction) on massive collections of unlabelled sequences, enabling the encoder to capture generic temporal patterns (Abbaspourazad et al., 2024). The pretrained representations are then transferred to diverse downstream applications with minimal task-specific adaptation. In addition, existing foundation models (Goswami et al., 2024; Gao et al., 2024; Pan et al., 2024) trained for time series data also adapt to large language models (LLMs). However, existing foundation models struggle to achieve accurate performance on irregular time series tasks, as shown in Figure 5.2, as they lack a mechanism to deal with the complexity of irregular sampling. Large pretrained models for irregular biosignal time series currently remain largely underexplored.

To this end, we introduce FORM[1], the first foundation model specifically designed for irregular biosignal time series. It is pretrained on diverse, multi-domain datasets spanning wearable biosignals and clinical physiological data, and is capable of generalising effectively across a wide range of downstream tasks and datasets. First, during pretraining, our architecture is designed to be variable-independent, enabling learning across multiple domains regardless of variable count or sequence length. To explicitly capture the inherent temporal dynamics and irregularities of irregular time series data, we train the variable-independent encoder with an irregular-sensitive mechanism via a masked reconstruction objective. Together, the pretraining stage learns dataset-agnostic representations enriched

[1]The work presented in this chapter is currently under review.

with a broad range of temporal dynamics and irregular patterns. During fine-tuning, these representations are refined through a specialised task-specific module equipped with multivariable attention mechanisms, capturing intricate intervariable dependencies and context-specific irregularities, thus enhancing performance on specific downstream tasks.

To support the practical evaluation of FORM, we conducted experiments on three large datasets of irregular and asynchronous time series that occur naturally in healthcare. Our contributions are summarised as follows:

- We highlight the significance and challenges posed by irregular biosignal time series in real-world applications, demonstrating substantial performance degradation of existing time series foundation models when applied to irregular data.
- We propose FORM, a pioneering large-scale pretraining framework uniquely designed for irregular biosignal time series, capable of learning generalisable representations that are robust to varying irregularities and easily adaptable to new domains with diverse variables and missingness patterns.
- FORM, pretrained on two large ICU datasets comprising more than 40,000 patients and 120 variables, demonstrates state-of-the-art performance and remarkable generalisation across three diverse biosignal datasets. Specifically, FORM surpasses baseline state-of-the-art irregular models by an average of 6.48% in irregular time series regression and by 3.5% in irregular time series classification tasks, showcasing its robustness and adaptability to different irregular scenarios.

5.1 Related work

In Section 2.4, we discuss the generalisation potential of existing foundation models but show that they still struggle with biosignal time series due to missingness and irregular sampling. This section reviews existing methods developed to alleviate irregularity, outlines their limitations, and provides a detailed discussion of why several foundation models fail to address the full complexity of such data.

5.1.1 Irregular time series modelling

Irregular time series are characterised by varying time intervals between adjacent observations (Shukla and Marlin, 2020). Early methods rely on set-based approaches, incorporating the time index as an additional feature, and using recurrent networks to learn irregular temporal dynamics (Che et al., 2016; Schirmer et al., 2022). Another line of work uses ODEs by parameterising the governing function in ODEs with neural networks. Furthermore, these methods combine ODEs with recurrent structures to learn the underlying dynamics

of time series, which intrinsically addresses irregularity (Chen et al., 2018; Rubanova et al., 2019). The attention mechanism in transform has also been improved to process irregular time series (Vaswani et al., 2017). For example, mTand (Shukla and Marlin, 2021) first replaces the positional encoding with fixed continuous-time embedding and then maps the irregular input into regular latent space. To account for the interchannel dependencies in the multivariate time series, graph-based approaches (Zhang et al., 2022c, 2024a; Yalavarthi et al., 2024) have also been explored by treating each variable as a node in the graph and learning the edge weights to capture the multivariate correlation. However, they are trained primarily under fully or semi-supervised paradigms (Tipirneni and Reddy, 2022; Zhang et al., 2023b), requiring large amounts of high-quality labelled or observed data. In contrast, approaches like PrimeNet (Chowdhury et al., 2023) and PATIS (Beebe-Wang et al., 2023) leverage self-supervised settings to learn representations of irregular time series and improve downstream tasks with limited labelled data, which outperform current supervised irregular methods such as ODEs. However, all existing methods design domain-specific model structures tailored for specific datasets by applying channel-dependent structures directly to the input to learn the correlations between multivariates, potentially hindering their adaptability to diverse datasets with different numbers of variables and tasks.

5.1.2 Foundation model for time series modelling

Foundation models, such as language models (Radford et al., 2019), are trained on broad data at scale to address diverse tasks with no or minimal additional training. Recent studies in time-series analysis have sought to develop models with similar capabilities. This includes developing novel architectures to capture diverse time series signals. For instance, Moment (Goswami et al., 2024) was pre-trained on multiple regular datasets using the patch and masked time-series reconstruction to learn general representations. UNITS (Gao et al., 2024) is a universal time series model that learns structural semantics from multi-domain data by pretraining a modified transformer. There have been several efforts to reprogram LLMs for time series tasks (Jin et al., 2024; Liu et al., 2025; Zhou et al., 2023). Models such as Time-LLM (Jin et al., 2024) and S2IP (Pan et al., 2024) adapt LLMs by fine-tuning their embedding layers or aligning time series samples with LLM-based text prototype embeddings. However, these models have so far been developed and evaluated only on regularly-sampled time series, and they lack the capability to model complex irregular time-series patterns. In our experiments (Section 5.5), we include several representative models (*e.g.*, MOMENT and CALF) as baselines and observe a significant performance drop when applied to irregular data, demonstrating their limited generalisability to such settings.

In contrast, our work seeks to fill this gap by designing a generalised large-scale

pretrained model using an SSL pretext task for irregular biosignal time series from multiple domains, not only effectively capturing the intrinsic characteristics of irregular time series but also learning more robust and generalisable representations.

5.2 Preliminaries

Our methods can deal with multiple irregular datasets, where each domain contains *multivariate irregular time series*. Samples can contain variable observation time lengths and numbers of sensors/variables, and the sampling time is irregular. Each dataset is denoted as $\mathbf{O} = (O^1, O^2, ..., O^D)$, where O^i represents the i^{th} variable among D variables. Each O^i is a *univariate irregular time series*, or a specific channel/variable, represented as $O^i = (o^i_{t_1}, o^i_{t_2}, ..., o^i_{t_{N_i}})$, recorded at time stamp $t_n, n \in [1, N_i]$. Unlike regular time series, the time interval between two consecutive measurements $\Delta t = t_{n+1} - t_n$ is not constant. The sequence length of observations N_i may also vary between different variables. We pad the sequence length to 512 by upsampling the original length, thereby ensuring a consistent number for each variable. To represent the irregularity, we simultaneously introduce an additional mask vector M and a time vector T to account for irregularities. Specifically, each dataset can be represented as $\mathbf{O} = (O^1, O^2, ..., O^D) = (\mathbf{T}, \mathbf{X}, \mathbf{M})$ where: $\mathbf{T} \in \mathbb{R}^{N \times D}$ represents the union of time stamps at which any of the D variables have been sampled. N denotes the total number of unique time stamps in the series. $\mathbf{X} \in \mathbb{R}^{N \times D}$ constitutes the observation values at the recorded times. $\mathbf{M} \in \{0, 1\}^{N \times D}$ is a binary masking matrix indicating whether a variable has an observation at a specific sampled time. Specifically, for each variable i at time t_n, if the variable is sampled at that time stamp, $M^i_{t_n} = 1$, otherwise $M^i_{t_n} = 0$.

5.3 Methods

5.3.1 FORM model overview

FORM is a multi-stage large pretrained framework designed to robustly model irregular time series data. It comprises two key components: (i) a *variable-agnostic pretraining* phase that learns from multiple datasets and domains by modelling each input channel independently; and (ii) a *task-specific fine-tuning* phase tailored to adapt the learned representations for specific downstream tasks and domains, incorporating inter-variable correlations. This design enables FORM to handle the irregularities inherent in real-world time series while generalising across diverse applications. An overview of FORM is illustrated in Figure 5.3.

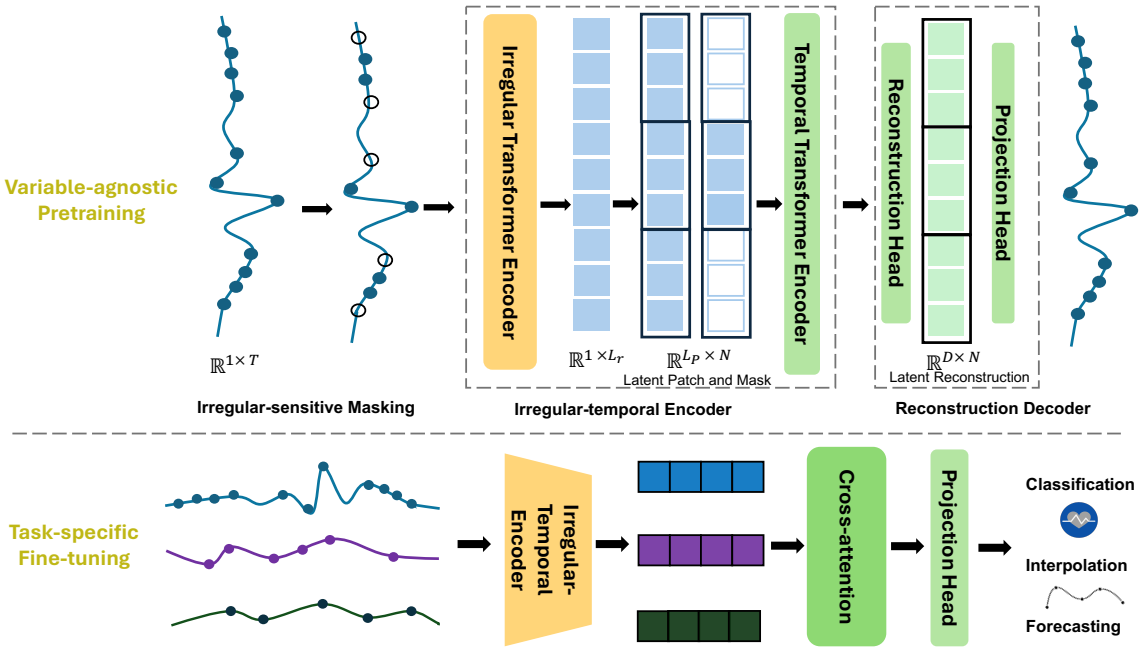


Figure 5.3: **Workflow of FORM:** a multi-stage pretrained framework designed for irregular time series modelling. (Top) During the variable-agnostic pretraining phase, each input channel is independently modelled through irregular-sensitive masking and reconstruction, enabling the encoder to learn robust temporal representations from diverse domains. (Bottom) During task-specific fine-tuning, the pretrained representations are adapted to downstream tasks by explicitly modelling inter-channel correlations using a cross-attention mechanism, effectively generalising to classification, interpolation, and forecasting tasks.

5.3.2 Variable-agonistic pretraining

The primary objective of the pretraining phase is to effectively capture the underlying temporal dynamics and irregular patterns present in multi-domain time series data using a unified model architecture. Unlike conventional methods that rely on dataset-specific training, FORM adopts a variable-independent approach: each variable (or channel) in each dataset is treated as a separate univariate sequence. This structure enables scaleable pretraining across heterogeneous datasets with varying numbers of variables and sampling characteristics. We adopt a reconstruction-based learning objective in which the model learns to reconstruct masked observations from partially observed sequences. This approach naturally fits the irregular setting, allowing the model to learn from both observed and missing patterns and adapt to various sampling gaps. Specifically, the FORM pretraining model comprises three core modules: (i) an *irregular-sensitive masking* module that masks segments of each variable while retaining irregular sampling intervals; (ii) an *irregular-temporal encoder* that transforms the input into a regularly spaced embedding and captures temporal dependencies through a patch-based representation; and (iii) a *reconstruction decoder* that predicts both masked and unmasked values using a lightweight prediction head. Together, these components enable the model to learn generalisable, variable-agnostic representations from irregular and asynchronous time series data.

Irregular-sensitive masking. This module aims to maintain the irregular patterns in the original time series. Specifically, recognising that different regions of the time series may have varying sampling densities, we mask a fixed time-span q_n within each variable rather than a constant number of observations (Chowdhury et al., 2023). This approach ensures that, in densely sampled regions, more observations are masked, while, in sparsely sampled regions, fewer observations are masked. As each variable may have a different sampling frequency and observation gap, we apply the masking strategy independently for each univariate sequence O^i.

Irregular-Temporal encoding. After the masking stage, the original irregular pattern is preserved. However, the irregularity challenge, *i.e.*, the unevenly sampled observations, persists within each sequence. To enable the reconstruction model to capture such irregular data, we first utilise an irregular transformer encoder to map the irregular time series into regularly spaced latent representations. Furthermore, to learn the temporal dependencies among these latent embeddings and reduce processing complexity, we further segment the regular embeddings into patches and design a global temporal encoding mechanism to capture the temporal dynamics across multiple patches.

Irregular transformer encoder. To map irregular data into regular latent space, we employ continuous-time embeddings and irregular-time attention(Shukla and Marlin, 2021) to encode the irregular time series into regular space, aiming to capture the local temporal dynamics. This approach consists of two key components: continuous-time embeddings and

the irregular-time attention (ITA) mechanism. Continuous-time embeddings function as irregular-aware positional embeddings, which incorporate irregular time points into a fixed vector space by leveraging H embedding functions $\phi_h(t)$, each outputting a representation of size d_r. Specifically, the dimension d of the embedding H is defined as:

$$\phi_h(t)[d] = \begin{cases} \omega_{0h} \cdot t + \alpha_{0h}, & \text{if } d = 0 \\ \sin(\omega_{dh} \cdot t + \alpha_{dh}), & \text{if } 0 < d < d_r \end{cases}, \tag{5.1}$$

where ω_{dh}'s and α_{dh}'s are learnable parameters. The linear term, when $d = 0$, captures non-periodic patterns that evolve over time, and the periodic terms using a sinusoidal function capture periodicity among time series data. This encodes the original irregular time indices into the high-dimensional space.

The subsequent ITA mechanism aims to convert each irregular time series O^i into irregular-aware embeddings Z^i by aligning the irregular time embeddings $\phi_h(t)$ with regular reference points $\mathbf{r}$ using an attention mechanism. This can convert the irregular time embeddings into a regular space while preserving the original irregular time information. Specifically, ITA uses the regular reference time points $\mathbf{r}$ as queries $Q = \phi_h(\mathbf{r})\mathbf{W}_q$, the observed irregular time points T_i as keys $K = \phi_h(T_i)\mathbf{W}_k$, and the original irregular time series X_i as values $V = X_i$. The attention mechanism is then employed to obtain the regular-spaced embeddings Z^i as follows:

$$Z^i = \sum_{h=1}^{H} \left(\text{Softmax} \left(\frac{\phi_h(\mathbf{r})\mathbf{W}_q[\phi_h(T_i)\mathbf{W}_k]^T}{\sqrt{d_k}} \right) X_i \right) \mathbf{W}_l \tag{5.2}$$

where $\mathbf{W_q}$, $\mathbf{W_k}$ are learnable parameter matrices, d_k is the dimension of the key vectors, and $\mathbf{W}_l$ is a learnable projection vector.

After this, the irregular sequences are transformed into regularly spaced latent representations. Since the irregular encoding maps arbitrary irregular time indices to a regular latent space while retaining irregularity-related information, it effectively learns and represents diverse irregular patterns. By maintaining these irregular characteristics, the model can generalise across different levels of irregularity, improving robustness and adaptability to various real-world datasets.

Temporal transformer encoder. To further capture the temporal dependencies within the sequence and benefit downstream tasks, we design a temporal transformer encoder to learn the global temporal dynamics. First, motivated by existing studies that utilise patches to enable more efficient processing, we further segment the regular-spaced embeddings into patches, $\mathbf{Z} = [Z_p^i]_{p=1}^P$ and apply random masks $\mathbf{M_i^{latent}}$ to make the encoder and the reconstruction model capture the temporal dependencies.

Specifically, we leverage the transformer multi-head attention mechanism to model

the temporal dependency within these patches, where $\mathbf{W} = \{\mathbf{W}_{\hat{q}}, \mathbf{W}_{\hat{k}}, \mathbf{W}_{\hat{v}}\}$, are learnable parameter matrices.

$$\text{MultiHead}(Z^i\mathbf{W}_{\hat{q}}, Z^i\mathbf{W}_{\hat{k}}, Z^i\mathbf{W}_{\hat{v}}) = [\text{head}_1, ..., \text{head}_h]\mathbf{W}, \tag{5.3}$$

$$\text{head}_h = \text{Softmax}(Z^i\mathbf{W}_{\hat{q}}[Z^i\mathbf{W}_{\hat{k}}]^T/\sqrt{d_k}) \tag{5.4}$$

The output of the multi-head attention mechanism will be fed into projection layers, as in a vanilla transformer encoder, to get our final embeddings $\boldsymbol{E}$ which capture both local irregular and global temporal dependency information.

Reconstruction loss and generalisation. The embeddings $\mathbf{E} = [E^i]_{i=1}^D$ learnt from the global temporal encoding will be further decoded by a decoder to reconstruct both the latent regular embedding $\boldsymbol{Z}$ and the original irregular time series $\hat{\mathbf{X}}$. The decoder is a simple MLP with layers designed to reconstruct the original time series at masked positions, chosen for their simplicity and computational efficiency. The reconstruction error between the model output $\hat{\mathbf{X}}$ and the target masked data $\mathbf{X}$ is computed using the mean squared error (MSE):

$$L = \frac{1}{D}(\lambda \sum_{i=1}^{D}(X^i - \hat{X}^i)^2 \cdot M^i + (1-\lambda)(Z^i - E^i)^2 \cdot M^i_{latent}) \tag{5.5}$$

5.3.3 Task-specific fine-tuning

After pretraining, the reconstruction model is fine-tuned to adapt the learnt representations for downstream tasks such as classification, forecasting, and interpolation on irregular multivariate time series. FORM supports flexible fine-tuning strategies: it can be trained end-to-end or used with linear probing by freezing all pretrained components and training only the task-specific head.

To better incorporate domain-specific information, FORM offers the variable-dependent fine-tuning mode. In the variable-independent pretraining setting, each variable is processed separately, preserving modularity and robustness to varying input dimensions. While this variable-agnostic design supports generalisation across heterogeneous datasets with differing sensor or channel numbers, it may limit the model's ability to capture cross-channel dependencies that are important for multimodal biosignal interpretation. In the variable-dependent fine-tuning stage, we introduce a cross-attention mechanism to explicitly model interactions between all variables, allowing the model to learn inter-variable relationships under task-specific supervision. Specifically, as illustrated in Figure 5.3 (bottom), the downstream input is passed through the pretrained encoder to generate latent representations $\boldsymbol{Z}$ for each variable. These serve as both keys and values in the attention mechanism, while the queries are either derived from the same set (self-attention)

Table 5.1: Comparison of datasets by number of variables and average number of observations.

Description	PhysioNet	MIMIC	HAR
Variable	37	96	12
Missing rate	79.6%	89.1%	75%
Sampling interval	1 hour	1 hour	100 ms

or from an aggregated task-specific embedding (cross-attention variant), which is computed as:

$$\text{Attention}(Z_Q, Z_K, Z_V) = \text{softmax}\left(\frac{Z_Q Z_K^\top}{\sqrt{d_k}}\right) Z_V, \tag{5.6}$$

This mechanism allows FORM to avoid overfitting to domain-specific variable correlations during pretraining, enabling the model to dynamically attend to informative signals and capture inter-variable dependencies that are most relevant for the downstream task.

Finally, a projection head combines the attended variable representations into a unified feature vector tailored for each task. This design allows FORM to leverage both the generalised temporal knowledge learnt during pretraining and the specific intervariable dependencies required for downstream performance.

5.4 Experiment setup

5.4.1 Datasets

We train and evaluate FORM on various benchmark irregular time biosignal series datasets spanning clinical and wearable health domains as described in Chapter 2.

- **Physionet** (Silva et al., 2012) consists of 37 time series variables extracted from intensive care unit (ICU) records.
- **MIMIC** (Johnson et al., 2016b) consists of electronic health records for more than 23457 patients with 96 variables.
- **Human Activity (HAR)** (Vidulin and Krivec, 2010) has 12 time series variables consisting of irregularly measured 3D positional records from 6, 554 sequences.

Specifically, we use both PhysioNet and MIMIC for pretraining, and evaluate the model on each of them independently to assess in-domain performance. HAR is used exclusively for evaluation, providing an outside-domain scenario that tests FORM's ability to generalise to nonclinical wearable-sensing data. The statistical details of each dataset are presented in Table 5.1.

5.4.2 Experimental protocols

We evaluate FORM across three critical time series tasks, namely *classification* (*e.g.*, in-hospital mortality), *interpolation* (filling in missing observations), and *forecasting* (predicting future values). For classification, the pretrained encoder is frozen or fine-tuned and paired with a task-specific classifier. For interpolation, the model predicts missing values by leveraging the temporal dynamics learnt during pre-training, which is directly aligned with the masked reconstruction objective. For forecasting, the model is trained to predict future time steps from historical context.

Classification task: To align with the regression settings, we expand the classification evaluation to 12,000 samples from the PhysioNet dataset. For MIMIC-III, we follow the same selection protocol as in previous work (Zhang et al., 2024a) and use 23,457 samples. The classification task involves binary prediction of hospital mortality, where the model predicts patient survival outcomes based on the input of irregularly sampled time series from each dataset. For the human activity recognition task, we follow the previous work (Shukla and Marlin, 2021) and focus on classifying each time point in the sequence into one of the 11 types of activities.

Interpolation task: For the interpolation task on the PhysioNet dataset, our objective is to predict missing values within a given time series. Specifically, we employ 8000 data cases. To evaluate the performance of the model under different conditions, we vary the number of observed points used for the prediction. Specifically, we test scenarios where 50% to 90% of the available data points are used to predict the remaining values during the test phase.

Forecasting task: For the forecasting task, we aim to use the previous data to predict the subsequent future observation values across various variables. Specifically, during testing, for PhysioNet and MIMIC, we use the first 24 hours to predict the next 24 hours. For Human Activity, we use the first 3000 milliseconds to predict the next 1000 milliseconds.

5.4.3 Baselines and implementation details

We compare FORM with state-of-the-art irregular time series models and foundation models for time series across various tasks. These methods are categorised into three groups: *(i) supervised models for irregular time series,* including ODE-based irregular time-series modelling, latent-ODE (Chen et al., 2018); Attention-based methods designed for irregular time series, including mTAND (Shukla and Marlin, 2021) and WarpFormer (Zhang et al., 2023b); Graph-based methods such as Raindrop (Zhang et al., 2022c) and tPatchGNN (Zhang

et al., 2024a); *(ii) self-supervised learning and foundation models for time series*, including PrimeNet (Chowdhury et al., 2023), and MOMENT (Goswami et al., 2024) ; *(iii) LLM-based foundation models for time-series*, including CALF (Liu et al., 2025). The details are described as follows:

- **Latent-ODE:** (Rubanova et al., 2019) A latent ordinary differential equation model that extends recurrent networks into continuous time. It defines hidden state dynamics with neural ODEs, enabling the model to capture the continuous-time evolution of time series, which is especially useful for irregularly sampled data.

- **mTAND:** (Shukla and Marlin, 2021) An attention-based architecture designed for irregularly sampled time series. It replaces discrete positional encoding with continuous-time embeddings, allowing the model to learn temporal relationships directly from the time gaps and align irregular observations onto a uniform representation space.

- **Warpformer:** (Zhang et al., 2023b) A transformer-based model tailored to irregular time series. It employs a specialised input encoding to represent both within-series irregular intervals and differences across series, and introduces a warping module that aligns sequences to a common time scale. Together with a custom attention mechanism, this design enables effective representation learning from unevenly sampled data.

- **T-patchGNN:** (Zhang et al., 2024a) A model that combines temporal convolution and graph neural networks to handle irregular time series. It first segments (or "patches") the irregular time series into uniform-length intervals and uses a time-aware convolutional network to encode these patches into latent features. A graph neural network is then applied to capture correlations between different variables (channels), improving multivariate forecasting on irregular data.

- **PrimeNet:** (Chowdhury et al., 2023) A contrastive learning framework for irregular time series data. PrimeNet learns time-sensitive representations by jointly optimising a contrastive objective (to differentiate time series instances based on their temporal patterns) and a reconstruction objective (to preserve the original signal). This approach encourages the model to capture irregularity patterns in the data, yielding robust embeddings that can be utilised for classification or interpolation tasks.

- **Moment:** (Goswami et al., 2024) An open-source family of large-scale pre-trained models (foundation models) for general-purpose time series analysis. MOMENT uses a high-capacity Transformer architecture and is pre-trained on a massive, domain-diverse dataset via a masked time-series prediction objective. This strategy yields

broad temporal representations, making MOMENT models versatile baselines that perform well across forecasting, classification, anomaly detection, and imputation tasks with minimal fine-tuning.

- **CALF:** (Liu et al., 2025) A cross-modal framework that aligns time series data with large language models for forecasting. It uses a dual-branch architecture: one branch processes multivariate time series directly, while the other converts the time series into a textual sequence fed into a pre-trained LLM. Through dedicated cross-modal alignment techniques (reducing distribution gaps and enforcing consistency between the temporal and textual paths), CALF enables the LLM to effectively model temporal patterns. This approach achieves state-of-the-art forecasting accuracy and demonstrates strong generalisation in few-shot and zero-shot settings.

5.4.4 Implementation details

Our implementation consists of two main stages: representation learning (pretraining) and task-specific fine-tuning. In the representation learning stage, we pretrain the model using a masked reconstruction objective to capture temporal patterns and irregularities, enabling robust and transferable representations across downstream tasks. For each dataset and task, we split the data into training, validation, and test sets using a 6:2:2 ratio. The representation model is pretrained using the full training set, then fine-tuned for the specific task by attaching task-specific projection heads based on the dataset requirements. Final evaluations are performed on the test set. To optimise performance, we conduct a comprehensive grid search over key hyperparameters. The batch size is varied among [32, 64, 128], while the learning rate is tested at [0.01, 0.001, 0.0001]. For the model architecture, we explore different configurations: The number of heads (H) in the irregular-transformer and cross-attention modules is varied among [1, 2, 3], and the number of transformer encoder layers is tested at [2, 4, 6]. To preserve irregularity within each univariate time series, we implement a masking strategy controlled by the timespan parameter q_n. This parameter, varied among [2, 3, 4], ensures that the masking ratio for each univariate series remains between 30% and 50%. The choice of patch size, critical for handling data sparsity, is selected from [32, 64, 128] based on the temporal density characteristics of the data. All experiments are conducted 5 times and the results are reported with mean and standard deviation. All models are implemented using PyTorch, and the experimental evaluations are conducted on NVIDIA A100-SXM-80GB GPUs.

5.4.5 Metrics

For all tasks, we select hyper-parameters on the held-out validation set using grid search and then apply the best-trained model to the test set. We randomly divide the dataset

into training, validation, and test sets using ratios of 60%, 20%, and 20% the same as previous studies (Shukla and Marlin, 2021; Zhang et al., 2024a). To assess the prediction of continuous targets, we use common regression metrics, such as the mean absolute error (MAE), mean-squared error (MSE) and root-mean-squared error (RMSE).

5.5 Results

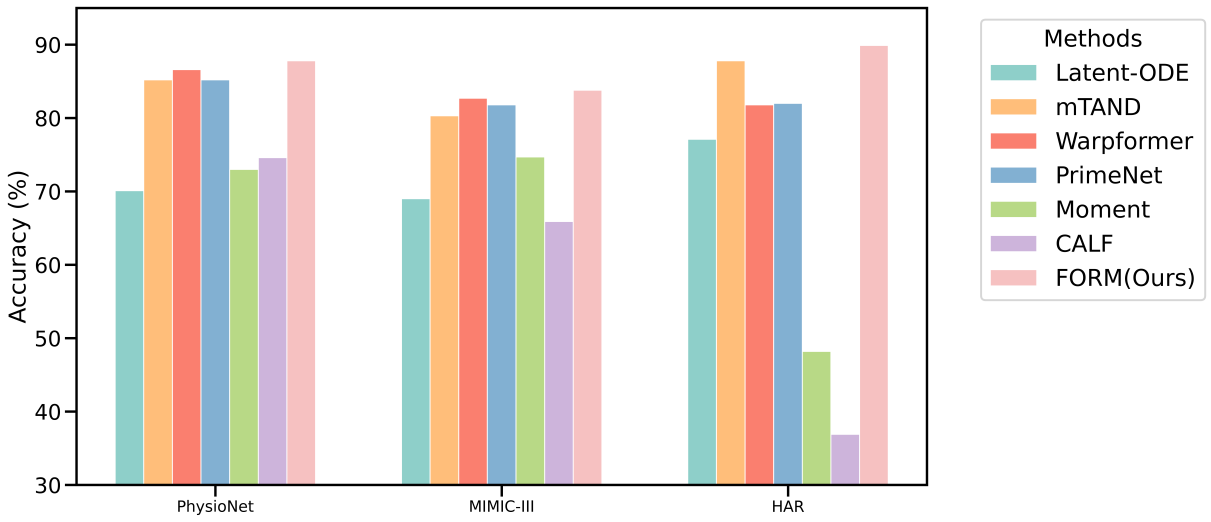


Figure 5.4: **Classification Performance across 3 Benchmark Datasets:** The figure shows that FORM achieves consistently higher accuracy than all baselines, demonstrating strong generalisation to diverse irregular time series and outperforming both irregular-TS methods and regular-TS foundation models.

5.5.1 Irregular biosignal time series modelling

Classification task. For self-supervised and LLM-based models (*e.g.*, Moment), we directly fine-tune the pretrained weights to perform the downstream classification tasks. For the other baseline methods, we strictly adhere to their original implementations, training them in an end-to-end supervised manner. Figure 5.4 provides a clear visual comparison of classification performance on three benchmarks. Our proposed method consistently achieves superior performance across all datasets, demonstrating its ability to learn distinct and robust representations tailored to each dataset. Specifically, it outperforms existing irregular time series methods such as Warpformer and PrimeNet, with average improvements of approximately 2.0% and 3.5%, respectively. Notably, foundation models originally designed for regular time series (*e.g.*, MOMENT and CALF) show significantly poorer performance compared to models specifically designed for irregular data. This performance advantage comes from FORM's multi-domain pretraining, which exposes the model to a broader range of irregular patterns and missingness conditions. This allows it

to better capture generalisable temporal dynamics and effectively transfer knowledge to downstream tasks, even in high-missingness settings such as MIMIC.

Table 5.2: **Forecasting performance** on four real-world irregular datasets evaluated using MSE and MAE, with best and second-best results in **bold** and underlined. FORM consistently achieves the lowest errors across all datasets.

Model	PhysioNet		MIMIC		Human Activity	
	MSE×10^{-3}	MAE×10^{-2}	MSE×10^{-2}	MAE×10^{-2}	MSE×10^{-3}	MAE×10^{-2}
Latent-ODE	6.05 ± 0.57	4.23 ± 0.26	1.89 ± 0.19	8.11 ± 0.52	3.34 ± 0.11	3.94 ± 0.12
mTAND	6.23 ± 0.24	4.51 ± 0.17	1.85 ± 0.06	7.73 ± 0.13	3.22 ± 0.07	3.81 ± 0.07
Warpformer	5.94 ± 0.35	4.21 ± 0.12	1.73 ± 0.04	7.58 ± 0.13	2.79 ± 0.04	3.39 ± 0.03
tpatchGNN	4.98 ± 0.08	3.72 ± 0.03	1.69 ± 0.03	7.22 ± 0.09	2.66 ± 0.03	3.15 ± 0.02
PrimeNet	5.33 ± 0.62	5.31 ± 0.68	1.87 ± 0.05	9.03 ± 0.29	2.94 ± 0.04	3.56 ± 0.07
Moment	142.0 ± 0.01	30.6 ± 0.02	10.4 ± 0.05	22.15 ± 0.13	252.0 ± 0.02	48.64 ± 0.32
CALF	67.0 ± 0.01	17.3 ± 0.06	8.30 ± 0.01	19.7 ± 0.01	130.0 ± 0.01	33.5 ± 0.01
Ours	**4.66 ± 0.07**	**3.38 ± 0.03**	**1.67 ± 0.10**	**7.03 ± 0.11**	**2.64 ± 0.05**	**3.13 ± 0.02**

Table 5.3: **Interpolation performance** on the **PhysioNet** dataset for varying percentages of observed data. FORM maintains the lowest error across all missingness levels.

Model	**Mean Squared Error** ($\times 10^{-3}$)				
Impute %	**50%**	**60%**	**70%**	**80%**	**90%**
Latent-ODE	6.72 ± 0.11	6.81 ± 0.05	6.80 ± 0.14	6.86 ± 0.06	7.14 ± 0.06
mTAND	4.14 ± 0.03	4.02 ± 0.05	4.16 ± 0.04	4.41 ± 0.15	4.80 ± 0.04
Warpformer	4.08 ± 0.05	4.05 ± 0.13	4.13 ± 0.06	4.46 ± 0.09	4.89 ± 0.17
tpatchGNN	3.83 ± 0.23	3.81 ± 0.16	3.85 ± 0.25	3.97 ± 0.33	3.99 ± 0.13
PrimeNet	3.28 ± 0.17	3.36 ± 0.12	3.42 ± 0.11	3.56 ± 0.11	3.66 ± 0.11
Moment	135.0 ± 0.01	137.0 ± 0.02	139.0 ± 0.02	141.0 ± 0.01	143.0 ± 0.02
CALF	97.6 ± 0.18	98.2 ± 0.08	95.4 ± 0.15	100.4 ± 0.03	101.6±0.05
Ours	**2.83 ± 0.13**	**2.87 ± 0.16**	**3.06 ± 0.07**	**3.13 ± 0.16**	**3.28 ± 0.19**

Forecasting task. Table 5.2 presents the forecasting performance of various methods across three benchmark datasets. FORM consistently outperforms all baselines in both MSE and MAE metrics in all data sets, highlighting its effectiveness in modelling temporal dependencies in irregular time series. Specifically, FORM achieves an average improvement of 4.33% in MSE and 6.48% in MAE compared to the second best method (typically tPatchGNN). These results reinforce two key findings: (1) self-supervised learning (SSL)-based methods outperform traditional end-to-end baselines by leveraging broader structural

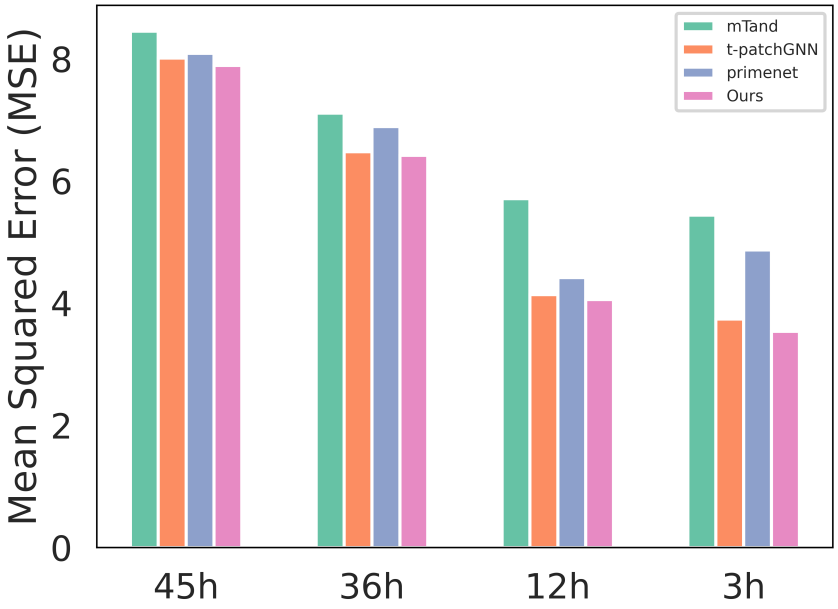


Figure 5.5: **Forecasting performance with various prediction horizons.** The x-axis denotes the prediction horizon. As longer observed segments are provided, forecasting performance steadily improves. Across all horizons, our method consistently surpasses baseline models, highlighting its robust and stable forecasting ability.

priors, and (2) most foundation models developed for regular time series, such as MOMENT and CALF, fail to generalise to the irregular setting, underscoring the need for irregularity-aware design. This performance gap is rooted in the inherent assumptions of these models. MOMENT and CALF are pretrained extensively on large-scale, regularly sampled time series data, where uniform time intervals are implicitly encoded in the architecture and training objectives. When applied to irregular data, these models rely on naive grid-based re-alignment strategies, which fail to preserve meaningful timestamp information and distort temporal patterns. As a result, they exhibit substantial degradation in forecasting performance, particularly in regression tasks involving high levels of missingness or nonuniform sampling intervals. To further validate robustness, we evaluate forecasting performance under varying observed lengths and forecast horizons, where the model uses previously observed time segments to predict future unseen intervals. Specifically, we show that FORM consistently outperforms SOTA methods when predicting different horizons on the PhysioNet dataset in Figure 5.5. The consistent superior performance across all time horizons underscores FORM's versatility and robustness in handling various forecasting scenarios, making it a reliable choice for a wide range of temporal prediction tasks.

Interpolation task. Table 5.3 shows the performance of FORM in the interpolation task using the PhysioNet dataset under varying masking ratios (from 50% to 90%). Each

masking ratio reflects the proportion of observed values removed during the testing, simulating different levels of missingness. FORM consistently outperforms all baselines across all masking levels, achieving the lowest mean squared error (MSE) in each case. Notably, it achieves a minimum MSE of 2.83×10^{-3} at 50% masking, representing an average relative improvement of 13.7% to 22.4% compared to the best-performing baselines such as PrimeNet and tPatchGNN. This demonstrates the model's robustness and adaptability under increasing levels of data sparsity. Additionally, large pretrained models such as MOMENT suffer significant performance degradation, indicating their limitations in handling irregular patterns when not explicitly trained for such data. These results highlight the value of variable-agnostic pretraining and irregular-sensitive encoding in enabling effective interpolation in real-world, sparse, irregular time series.

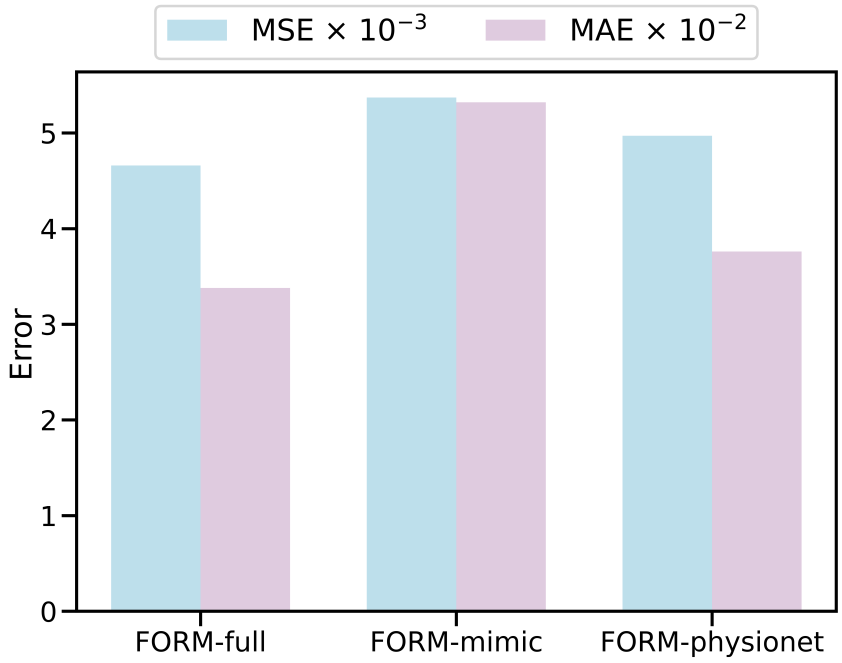


Figure 5.6: **Generalisability of FORM fine-tuned on the PhysioNet forecasting task.** We compare models pretrained on different scenario with various datasets: FORM-full, FORM-mimic and FORM-physionet. FORM pretrained on more diverse datasets (FORM-full) achieve the lowest errors, showing that multi-domain pretraining provides stronger transferable temporal priors than single-domain pretraining.

5.5.2 Analysis of generalisation capabilities

To further assess the generalisability of FORM, we performed a cross-domain transfer learning experiment using the PhysioNet forecast task as the target domain. Given FORM's variable-agnostic pretraining design, we compare its performance when pretrained under three different scenarios: (1) FORM-full: pretrained on both MIMIC and PhysioNet (96 + 37 variables), (2) FORM-mimic: pretrained only on MIMIC (96 variables), (3)

FORM-physionet: pretrained only on PhysioNet (37 variables). This setup enables us to evaluate the impact of pretraining domain diversity on model generalisability. As shown in Figure 5.6, FORM-full achieves the best performance in both the MSE and MAE metrics, benefiting from exposure to a wider range of temporal patterns and irregularities. Interestingly, FORM-physionet outperforms FORM-mimic, despite being pretrained on a smaller dataset, indicating that pretraining on the same domain still provides a stronger inductive prior than cross-domain pretraining alone. These results highlight two insights: first, multi-domain pretraining enables stronger generalisation than any single-domain model; and second, variable-agnostic pretraining helps capture transferable temporal dynamics without overfitting to inter-variable correlations. This validates FORM's design as an effective and generalisable approach to modelling irregular time series across diverse domains.

Table 5.4: **Ablation study.** Removing any component of FORM leads to clear performance degradation and both the irregular-sensitive masking and global embedding modules are essential for capturing irregular pattern and temporal dependency.

Model	PhysioNet		MIMIC		Human Activity	
	MSE$\times 10^{-3}$	MAE$\times 10^{-2}$	MSE$\times 10^{-2}$	MAE$\times 10^{-2}$	MSE$\times 10^{-3}$	MAE$\times 10^{-2}$
FORM (Full)	**4.66 ± 0.07**	**3.38 ± 0.03**	**1.67 ± 0.10**	**7.03 ± 0.11**	**2.64 ± 0.05**	**3.13 ± 0.02**
MAE (**w/o all**)	7.18 ± 0.18	4.95 ± 0.13	3.95 ± 0.16	12.82 ± 0.21	4.34 ± 0.11	4.94 ± 0.12
w/o IR-mask	6.81 ± 0.28	4.76 ± 0.19	2.95 ± 0.07	9.85 ± 0.11	3.34 ± 0.38	3.94 ± 0.12
w/o GE	5.86 ± 0.45	4.45 ± 0.53	2.63 ± 0.47	9.13 ± 0.33	3.98 ± 0.48	5.12 ± 0.34

5.5.3 Ablation study

We evaluated the performance of FORM and its variants on all datasets used for the regression task. **Complete** represents the model without any ablation; **w/o Ir-mask** removes the irregular-sensitive masking strategy while using the random masking strategy; **w/o GE** (global embedding) removes the temporal and global embedding module, including latent masking and patching and **w/o all** is just the simple masked autoencoder framework. Table 5.4 presents the results of the model ablation study. As shown, the removal of any component leads to performance degradation. Notably, the **w/o IR-mask** configuration enables a significant performance drop in all datasets, demonstrating the importance of capturing and maintaining the original irregularity when using the masking and reconstruction mechanism. Specifically, for Human Activity with denser regions and MIMIC-III with sparser regions, similar improvements of 34% and 28% are observed. This indicates that our method can capture various irregular patterns and has the capability to generalise to different domains. Additionally, the absence of a transformer for learning global temporal correlations further emphasises the importance of modelling long-range

dependencies in irregular time series. Without such a mechanism, the model struggles to capture overarching temporal structure, leading to less informative representations and reduced forecasting accuracy. This highlights the necessity of explicitly incorporating global temporal correlations during representation learning to enable more effective regression forecasting.

5.6 Discussion and conclusions

This chapter builds upon the ideas and limitations identified in the previous chapters to develop a general-purpose foundation model for biosignal time series. Building on Chapter 3's exploration of self-supervised pretraining on large unlabelled data and Chapter 4's work on mitigating real-world label distribution shifts, this chapter addresses the more fundamental challenge of modelling complex biosignal data that exhibit irregular sampling, missingness, and cross-dataset variability.

Existing biosignal models typically perform well only on regularly sampled data or require dedicated training models tailored specifically to individual datasets with fixed numbers of variables or specific missingness patterns. To overcome these limitations, we introduce FORM, a generalised pretraining framework for irregular biosignal time series. By incorporating irregular-sensitive masking, an irregular-time encoder, and global temporal encoding, FORM effectively captures the underlying dynamics for representation learning. Its channel-independent design enables strong performance across domains, establishing it as a foundational model for diverse irregular time series tasks. This work offers a robust and generalisable framework for real-world applications.

Through extensive evaluation on benchmark healthcare datasets, we demonstrated that FORM achieves state-of-the-art predictive performance, is robust to irregular sampling, and exhibits strong generalisation across tasks and datasets, spanning both classification and regression. This chapter thus provides a scaleable and generalisable framework for modelling irregular biosignal data in real-world applications.

While FORM demonstrates strong performance, it still presents several limitations and avenues for improvement. First, its pretraining is computationally intensive due to variable-wise processing and fine-grained reconstruction. For example, although we now show that the model can learn effective representations for 2-day biosignals, scaling to longitudinal month- or year-level data from consumer wearables will likely require a more efficient and distributed training pipeline. Second, the fixed time-span masking used during pretraining is a simple heuristic and may not adapt to the heterogeneous sampling densities found in real-world biosignals. An adaptive masking strategy that learn mask lengths based on local sampling density or temporal uncertainty could more effectively capture the model to diverse irregularity patterns. Third, the current evaluation relies

primarily on low-frequency clinical and HAR datasets, which limits the assessment of FORM's broader generalisation to more diverse and longitudinal settings. Future work could explore more efficient training strategies, adaptive masking mechanisms, and broader and longitudinal data curation to address these challenges.

Chapter 6

Conclusions and future directions

At the beginning of this thesis, we emphasised the importance of designing effective and generalisable machine learning models for biosignal time series, with the ultimate goal of improving health monitoring and human well-being. Biosignal data, ranging from clinical physiological signals to wearable sensor measurements, offer immense potential for early detection, intervention, and personalised care. However, traditional methods and current technologies remain limited as they often assume regular sampling, require extensive labelled data, and struggle to generalise across datasets or tasks.

To address these limitations, this thesis investigated three complementary research directions: (i) self-supervised learning to leverage large-scale unlabelled biosignal data; (ii) domain adaptation to mitigate cross-domain distribution shifts in real-world deployment; and (iii) foundation models for irregular time series, culminating in a scaleable and general-purpose approach to biosignal modelling.

With regard to contributions, we presented three original pieces of work, each grounded in fundamental research questions about how to build effective and generalisable models for healthcare biosignals. Our central premise is that new training paradigms can create more robust, transferable, and trustworthy representations. These, in turn, can be applied to a wide range of healthcare scenarios, from continuous patient monitoring in clinical settings to daily well-being tracking with consumer wearables.

In this chapter, we briefly summarise the key contributions of this thesis, reflect on their implications for the broader field of digital health, and suggest directions for future research.

6.1 Summary of contributions

In this section, we reflect on the research questions introduced in Chapter 1 and summarise the major contributions of this thesis.

6.1.1 Self-supervised learning for unlabelled biosignal time series data

Research Question 1. *How can we use machine learning to learn representations from unlabelled biosignal time series data and perform effectively and achieve strong performance on diverse classification tasks?*

Contribution 1. In Chapter 3, we presented StatioCL, a novel CL framework that captures inherent non-stationarity and temporal dependency for large unlabelled biosignal time series representation learning. The motivation behind this study stemmed from the limitations of existing CL approaches. In the time series modelling domain, CL has emerged as a promising approach for representation learning by embedding similar pairs closely while distancing dissimilar ones. However, existing CL methods often introduce false negative pairs (FNPs) by neglecting inherent characteristics and then randomly selecting distinct segments as dissimilar pairs, leading to erroneous representation learning, reduced model performance, and overall inefficiency.

To address these issues, we systematically defined and categorised FNPs in time series into *semantic false negative pairs* and *temporal false negative pairs* for the first time: the former arising from overlooking similarities in label categories, which correlate with similarities in non-stationarity and the latter from neglecting temporal proximity. By interpreting and differentiating non-stationary states, which reflect the correlation between trends or temporal dynamics with underlying data patterns, StatioCL effectively captures the semantic characteristics and eliminates semantic FNPs. Simultaneously, StatioCL establishes fine-grained similarity levels based on temporal dependencies to capture varying temporal proximity between segments and mitigate temporal FNPs. We found a StatioCL demonstrates substantial improvement over state-of-the-art CL methods, achieving a 2.9% increase in recall and a 19.2% reduction in FNPs when evaluated on real-world benchmark time series classification datasets,. Most importantly, StatioCL also shows improved data efficiency and robustness against label scarcity.

6.1.2 Domain adaptation to alleviate noisy labels to improve healthcare prediction via free-living wearable devices

Research Question 2. *How can we generalise models trained on large-scale free-living biosignal data with noisy or weak labels to accurately predict gold-standard health outcomes?*

Contribution 2. In Chapter 4, we presented a multi-discriminator domain adaptation framework for cross-domain representation learning that alleviates the shift of the label distribution by transferring knowledge from weakly labelled large-scale data to smaller but

high-quality health datasets. The motivation behind this approach is that high-quality datasets with gold-standard labels are expensive and difficult to collect in the healthcare domain, leading to sparse and small-scale datasets that make it difficult to generalise to other unseen cohorts and applications. In contrast, collecting silver-standard labels from less accurate collection schemes with modern wearables is more affordable. However, these less accurate, extensive labels exhibit distribution discrepancies compared to ground-truth labels and cannot be directly leveraged for model deployment.

To address this challenge, our framework introduces fine-grained multi-discriminator adaptation, where multiple discriminators operate across different feature subspaces to more effectively align weak-label and gold-label domains. This design allows the model to account for complex discrepancies in label distributions and to leverage large-scale wearable data without degrading performance on clinical ground-truth targets.

We validated this approach on the cardio-fitness prediction task. Our results showed that leveraging large-scale noisy VO_2max labels using the proposed fine-grained multi-discriminator domain adaptation not only achieves improved fitness prediction but also effectively mitigates label distribution shifts. Specifically, our model outperforms competing methods by up to 12% compared to competitive transfer learning and state-of-the-art domain adaptation models. This contribution demonstrates that carefully designed domain adaptation methods can bridge the gap between scaleable but noisy wearable datasets and scarce but precise clinical datasets, paving the way for more practical and robust machine learning applications in healthcare.

6.1.3 Foundation model for multi-domain irregular time series

Research Question 3. *How can we build general-purpose foundation models that handle the complexities of biosignal time series data, including irregular sampling, missingness, and heterogeneity of the modality, while effectively generalising across domains and downstream healthcare tasks?*

Contribution 3. In Chapter 5, we developed a general-purpose foundation model that unifies insights from self-supervised learning and cross-domain generalisation to learn robust representations from irregular biosignal datasets. The goal was to go beyond dataset-specific approaches and create a model that could generalise across unseen tasks and domains through transfer learning.

To achieve this, we introduced FORM, a novel foundation model specifically designed for irregular biosignal time series, pretrained across diverse, multi-domain datasets. Our method employs a variable-independent encoder trained with a masked reconstruction objective, enabling the model to learn variable-agnostic representations that capture a broad range of temporal dynamics and irregularities. During fine-tuning, these representations

are processed through a task-specific module equipped with multi-variable attention mechanisms to model inter-variable dependencies and context-specific irregularities more effectively.

We evaluated FORM across multiple irregular biosignal datasets and tasks, including classification, forecasting, and interpolation, and it consistently demonstrated state-of-the-art performance. Beyond evaluation of the within-dataset, we also tested its adaptability by transferring the pre-trained model across domains: even when fine-tuned on unseen datasets with different variables and patterns of irregularity, FORM maintained competitive accuracy. This work presents a scaleable and generalisable framework for modelling complex, irregular time series data in real-world applications.

6.2 Implications and limitations

This thesis explores three directions in biosignal modelling to build effective and generalisable health predictors. Across these contributions, significant effort was devoted to addressing key data challenges such as missingness, irregularity, label scarcity, and cross-domain generalisation. The findings hold promising implications for multiple communities and stakeholders.

For the research community, this work contributes methodological insights into how biosignal-specific characteristics, such as non-stationarity, irregular sampling, and weak labelling, can be systematically addressed with modern machine learning paradigms. Researchers in biosignals and deep learning can build on these frameworks to design more advanced models, generate richer representations, and derive clinically meaningful inferences for their own datasets and applications. For engineers and industry practitioners, the proposed approaches suggest new opportunities for building advanced wearables and machine learning–driven health products. By leveraging large-scale unlabelled data and robust pretraining strategies, consumer devices can enable daily fitness and health monitoring without relying exclusively on clinically validated ground-truth labels. This opens the door to scaleable, low-cost monitoring solutions that provide users with more actionable insights into their health and well-being. For clinicians, this thesis highlights how passively collected person-generated data from daily life can complement episodic data obtained in routine clinical practice. Such integration can improve the efficiency and coverage of clinical diagnosis, for example, by providing continuous context between clinical visits or early warning signals for conditions that episodic assessments may miss. Ultimately, this work suggests a pathway for bridging daily health monitoring and clinical decision making, contributing to a more proactive, personalised, and timely delivery of healthcare.

It is important to acknowledge the limitations of this thesis so that its findings can

be interpreted in the correct context and future research can build on them. While the work presented here addresses core challenges in biosignal modelling such as missingness, irregularity, and domain shifts, several open issues remain.

Our evaluations primarily focused on the technical aspects of biosignal modelling rather than explicitly examining population-specific differences. However, biosignals and user behaviour can vary substantially between demographic, age groups, and cultural contexts and can also shift over time. For example, wearable heart rate data from young athletes differ in dynamics and noise characteristics compared to those from elderly patients in clinical settings. These variations raise concerns about fairness and generalisability between populations. Ensuring equitable performance in personalised health applications requires more systematic evaluation in diverse cohorts, as well as a fairness-based model design (Ahmad et al., 2020).

Although our research concentrated on biosignal time series, modern health data ecosystems are increasingly multimodal. In addition to wearable signals, data sources such as electronic health records (EHRs), clinical notes, and self-reported surveys provide valuable context. For example, supplementing noisy wearable sleep data with EHR-based diagnoses could improve predictive robustness (Rajkomar et al., 2018). Although our methods alleviate missingness within biosignals, they do not yet leverage knowledge from other modalities. Incorporating multimodal fusion to combine signals, text, and structured clinical data represents an important future direction for strengthening robustness and interpretability.

Our proposed foundation model for irregular time series (FORM) demonstrated strong performance, but scaling such models presents significant computational challenges. Biosignal data sets are vast, heterogeneous and collected with different numbers of sensors and recording durations. Training variable-independent encoders across such diverse sources is resource-intensive, and building truly large-scale pretrained biosignal foundation models remains costly in terms of both computation and memory. Moreover, while we demonstrated promising generalisation to new datasets and tasks, the degree of true universality remains an open question. Therefore, more efficient training paradigms, such as parameter-efficient tuning or architectures designed for efficiency, will be necessary for practical deployment.

Another important point worth mentioning is that while this thesis primarily focused on short-term prediction tasks (*e.g.*, VO_2max estimation or 48-hour in-hospital mortality), real-world healthcare often requires longitudinal and time-to-event modelling. Wearable devices uniquely enable continuous monitoring, capturing long-term health trends such as gradual decline in fitness or early onset of diseases (Nakamura et al., 2016). However, our work did not fully explore how biosignal models can support long-term outcome prediction or integrate predictions back into clinical workflows. Bridging this gap by connecting predictive models with actionable interventions, diagnosis support, and patient

participation will be critical to translating technical advances into a real-world health impact. Finally, we emphasise that there is no universal solution to the challenges of biosignal modelling. Missingness, irregularity, heterogeneity, and fairness are deeply intertwined with the temporal and causal nature of health data. However, by carefully formulating research questions and designing methods that respect temporal dynamics and causal structure, we believe that meaningful progress can continue to be made.

6.3 Future research directions

In this thesis, we have answered three research questions introduced in Chapter 1; nevertheless, many future directions were naturally uncovered.

6.3.1 Foundation models for biosignals: training from scratch or leveraging large language models for biosignal analysis

As discussed in Chapter 5, foundation models are large neural networks trained in diverse datasets, typically through self-supervised learning, and subsequently adaptable to a wide range of downstream tasks. In biosignal analysis, three complementary directions have emerged: (i)training foundation models from scratch on large biosignal datasets (Yuan et al., 2022; Zhang et al., 2023a); (ii)adapting general time series foundation models to biosignal domains (Goswami et al., 2024; Gao et al., 2024); and (iii)leveraging large language models (LLMs) as backbones or interactive agents (Yang et al., 2024).

Our work in this thesis primarily pursued the first direction, specifically training biosignal-specific foundation models from scratch to explicitly address irregular sampling and missingness. This approach demonstrated strong effectiveness, but also highlighted limitations, particularly in terms of computational cost and scalability. Future research should therefore explore how to balance the strengths of scratch-trained biosignal models with the transferability of general time series foundation models and the semantic reasoning capabilities of LLMs.

In particular, recent studies are increasingly investigating how LLMs can be integrated into biomedical time series analysis. LLMs excel in semantic reasoning, contextual understanding and generative capabilities, and recent advances extend these abilities to multimodal contexts (*e.g.*, images, audio, and time series) (Zhang et al., 2024b). Building on these advances, several promising future directions emerge. First, LLMs can be treated as powerful backbones for biosignal modelling. By reprogramming biosignals into token-based sequences or textual formats, pretrained LLMs can serve as direct modelling architectures (Ji et al., 2024). For instance, biosignal waveforms could be represented as structured textual descriptions ("heart rate variability: low") or transformed into

spectrograms for vision-language LLMs. These strategies would allow biosignal models to inherit LLMs' pretrained reasoning capacity with minimal additional training. Besides, LLMs can also act as domain-knowledge providers. LLMs can act as domain-informed teachers by enriching representations with detailed textual annotations (Yu et al., 2024), contextualising biosignal patterns with medical knowledge, or aligning signal features with semantic descriptions. For example, wearable-derived VO_2max estimates could be supplemented with textual explanations (*e.g.*, "indicative of low aerobic capacity"), improving supervision in low-label regimes. Finally, beyond static predictions, LLMs open the possibility of conversational agents and workflow-based agents that can interpret biosignals in real time, answer patient or clinician queries, and coordinate multistep analysis pipelines (Heydari et al., 2025). For example, a wearable-powered "biosignal copilot" could explain fluctuations in heart rate variability to users, contextualised with activity or stress levels, or offer clinicians structured, longitudinal summaries that synthesise trends, anomalies, and clinically meaningful patterns across extended periods of sensor data.

Together, these directions suggest that the future of biosignal foundation models may not lie in a single strategy but rather in hybrid approaches that combine the strengths of biosignal-specific pretraining, general TSFMs, and multimodal LLMs.

6.3.2 Test-Time adaptation for biosignals

In previous chapters, we highlighted that supervised models are typically trained under the assumption of full source-domain availability, without accounting for the realities of data influx and distribution shifts during the real-time testing phase. However, in practice, biosignal data collected at test time often differ significantly from training data due to environmental changes, sensor variability, population differences, or inconsistencies in labelling protocols. Although our proposed fine-grained representation learning methods and foundation models alleviate some of these issues by improving cross-domain robustness, they remain limited in that adaptation is performed offline and often requires labelled samples from the target domain.

A promising avenue for future research is test-time adaptation (TTA), where models adapt on-the-fly during inference using only unlabelled data from the target domain (Chen et al., 2023a). In particular, Online Test-Time Adaptation (OTTA) approaches continuously update pretrained models with incoming test samples, enabling robustness to evolving distributions (Jo et al., 2024). Such methods are especially relevant for healthcare applications that require real-time monitoring and intervention, where waiting for labelled data or offline retraining is infeasible. Several directions have emerged in the broader machine learning community. For example, entropy minimisation and self-training methods adjust model predictions by encouraging confident outputs on incoming unlabelled samples (Wang et al., 2021). Auxiliary self-supervised objectives, such as masked recon-

struction or rotation prediction, can be applied online to refine representations without labels, helping models maintain stable and reduce drift as testing conditions evolve.

Despite these advances, biosignals remain an underexplored testbed for TTA. Wearable and clinical biosignal data offer unique challenges: non-stationarity, irregular sampling, multisensor heterogeneity, that differ from the image and text domains where TTA methods are usually evaluated. For example, heart rate variability can shift dramatically due to stress, sleep, or exercise; EEG patterns can drift between sessions or electrode setups; and wearable devices differ in sampling rates and noise levels. Therefore, designing biosignal-specific OTTA benchmarks and methods would be a valuable next step.

We see particular promise in integrating OTTA with wearable systems (Wang et al., 2024a). A smartwatch or portable device could continuously adapt its health monitoring model to the daily fluctuations and device-specific noise of a user, without requiring manual relabelling. This would allow systems to generalise not only across unseen domains but also across individual users and contexts in real time. By capturing the essence of biosignal data regardless of population, device, or environment, TTA could play a key role in bridging the gap between controlled research settings and real-world deployment.

Bibliography

Abbaspourazad, S., Elachqar, O., Miller, A. C., Emrani, S., Nallasamy, U., and Shapiro, I. (2024). Large-scale training of foundation models for wearable biosignals. In *ICLR*.

Abut, F. and Akay, M. (2015). Machine learning and statistical methods for prediction of maximal oxygen uptake: Recent advances. *Medical Devices: Evidence and Research*.

Abut, F., Akay, M. F., and George, J. (2016). Developing new vo2max prediction models from maximal, submaximal and questionnaire variables using support vector machines combined with feature selection. *Computers in biology and medicine*, 79:182–192.

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al. (2023). Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Ahmad, M. A., Eckert, C., and Teredesai, A. (2018). Interpretable machine learning in healthcare. In *Proceedings of the 2018 ACM international conference on bioinformatics, computational biology, and health informatics*, pages 559–560.

Ahmad, M. A., Patel, A., Eckert, C., Kumar, V., and Teredesai, A. (2020). Fairness in machine learning for healthcare. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 3529–3530.

Al-Turjman, F. and Baali, I. (2022). Machine learning for wearable iot-based applications: A survey. *Transactions on Emerging Telecommunications Technologies*, 33(8):e3635.

Alanazi, A. (2022). Using machine learning for healthcare challenges and opportunities. *Informatics in Medicine Unlocked*, 30:100924.

Aldhoayan, M. and Aljubran, Y. (2023). Prediction of icu patients' deterioration using machine learning techniques. *Cureus*, 15.

Alghatani, K., Ammar, N., Rezgui, A., and Shaban-Nejad, A. (2020). Predicting intensive care unit length of stay and mortality using patient vital signs: Machine learning model development and validation. *JMIR Medical Informatics*, 9.

Anguita, D., Ghio, A., Oneto, L., Parra, X., and Reyes-Ortiz, J. L. (2013). A public domain dataset for human activity recognition using smartphones. In *The European Symposium on Artificial Neural Networks*.

Ansari, A. F., Stella, L., Turkmen, C., Zhang, X., Mercado, P., Shen, H., Shchur, O., Rangapuram, S. S., Arango, S. P., Kapoor, S., et al. (2024). Chronos: Learning the language of time series. *arXiv preprint arXiv:2403.07815*.

Arora, S., Khandeparkar, H., Khodak, M., Plevrakis, O., and Saunshi, N. (2019). A theoretical analysis of contrastive unsupervised representation learning. In *International Conference on Machine Learning*.

Auepanwiriyakul, C., Waibel, S., Songa, J., Bentley, P., and Faisal, A. A. (2020). Accuracy and acceptability of wearable motion tracking for inpatient monitoring using smartwatches. *Sensors (Basel, Switzerland)*, 20.

Baevski, A., Zhou, Y., Mohamed, A., and Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460.

Baig, M. M. and Gholamhosseini, H. (2013). Smart health monitoring systems: An overview of design and modeling. *Journal of Medical Systems*, 37:1–14.

Beebe-Wang, N., Ebrahimi, S., Yoon, J., Arik, S. Ö., and Pfister, T. (2023). Paits: Pretraining and augmentation for irregularly-sampled time series. *ArXiv*, abs/2308.13703.

Bolpagni, M., Pardini, S., Dianti, M., and Gabrielli, S. (2024). Personalized stress detection using biosignals from wearables: A scoping review. *Sensors (Basel, Switzerland)*, 24.

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., et al. (2021). On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Bousmalis, K., Trigeorgis, G., Silberman, N., Krishnan, D., and Erhan, D. (2016). Domain separation networks. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, page 343–351. Curran Associates Inc.

Brage, S., Brage, N., Franks, P., Ekelund, U., and Wareham, N. (2005). Brage s, brage n, franks pw, ekelund u, wareham nj. reliability and validity of the combined heart rate and movement sensor actiheart. eur j clin nutr 59, 561-570. *European journal of clinical nutrition*, 59:561–70.

Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., and Joulin, A. (2020). Unsupervised learning of visual features by contrasting cluster assignments. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Castaneda, D., Esparza, A., Ghamari, M., Soltanpur, C., and Nazeran, H. (2018). A review on wearable photoplethysmography sensors and their potential future applications in health care. *International journal of biosensors & bioelectronics*, 4(4):195.

Che, Z., Purushotham, S., Cho, K., Sontag, D. A., and Liu, Y. (2016). Recurrent neural networks for multivariate time series with missing values. *Scientific Reports*, 8.

Chen, L., Zhang, Y., Song, Y., Shan, Y., and Liu, L. (2023a). Improved test-time adaptation for domain generalization. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24172–24182.

Chen, R. J., Wang, J. J., Williamson, D. F., Chen, T. Y., Lipkova, J., Lu, M. Y., Sahai, S., and Mahmood, F. (2023b). Algorithmic fairness in artificial intelligence for medicine and healthcare. *Nature biomedical engineering*, 7(6):719–742.

Chen, R. T. Q., Rubanova, Y., Bettencourt, J., and Duvenaud, D. (2018). Neural ordinary differential equations. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, page 6572–6583, Red Hook, NY, USA. Curran Associates Inc.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. (2020). A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org.

Chen, Y., Ren, K., Wang, Y., Fang, Y., Sun, W., and Li, D. (2023c). Contiformer: continuous-time transformer for irregular time series modeling. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Chowdhury, R. R., Li, J., Zhang, X., Hong, D., Gupta, R. K., and Shang, J. (2023). Primenet: Pre-training for irregular multivariate time series. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(6):7184–7192.

Chung, J., Çaglar Gülçehre, Cho, K., and Bengio, Y. (2014). Empirical evaluation of gated recurrent neural networks on sequence modeling. *ArXiv*, abs/1412.3555.

Dauwels, J., Vialatte, F., Musha, T., and Cichocki, A. (2010). A comparative study of synchrony measures for the early diagnosis of alzheimer's disease based on eeg. *NeuroImage*, 49(1):668–693.

Du, Y., Tan, Z., Chen, Q., Zhang, X., Yao, Y., and Wang, C.-J. (2020). Dual adversarial domain adaptation. *ArXiv*, abs/2001.00153.

Du, Z., Li, J., Su, H., Zhu, L., and Lu, K. (2021). Cross-domain gradient discrepancy minimization for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3937–3946.

Eldele, E., Ragab, M., Chen, Z., Wu, M., Kwoh, C. K., Li, X., and Guan, C. (2021). Time-series representation learning via temporal and contextual contrasting. *CoRR*, abs/2106.14112.

Esco, M., Mugu, E., Williford, H., Mchugh, A., and Bloomquist, B. (2011). Cross-validation of the polar fitness testtm via the polar f11 heart rate monitor in predicting vo2max. *Journal of Exercise Physiology*, 14:43–52.

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J. M., Swetter, S. M., Blau, H. M., and Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542:115–118.

Farahani, A., Voghoei, S., Rasheed, K., and Arabnia, H. R. (2021). A brief review of domain adaptation. *Advances in data science and information engineering: proceedings from ICDATA 2020 and IKE 2020*, pages 877–894.

Faurholt-Jepsen, M., Brage, S., Kessing, L. V., and Munkholm, K. (2017). State-related differences in heart rate variability in bipolar disorder. *Journal of psychiatric research*, 84:169–173. 27743529[pmid].

Ganapathy, N., Swaminathan, R., and Deserno, T. M. (2018). Deep learning on 1-d biosignals: a taxonomy-based survey. *Yearbook of Medical Informatics*, 27:98 – 109.

Ganin, Y. and Lempitsky, V. (2015). Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, pages 1180–1189. PMLR.

Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., and Lempitsky, V. (2016). Domain-adversarial training of neural networks. *J. Mach. Learn. Res.*, 17(1):2096–2030.

Gao, S., Koker, T., Queen, O., Hartvigsen, T., Tsiligkaridis, T., and Zitnik, M. (2024). Units: a unified multi-task time series model. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA. Curran Associates Inc.

Gehring, J., Auli, M., Grangier, D., Yarats, D., and Dauphin, Y. (2017). Convolutional sequence to sequence learning. *ArXiv*, abs/1705.03122.

Ghulam, M., Alshehri, F., Karray, F., Saddik, A. E., Alsulaiman, M., and Falk, T. H. (2021). A comprehensive survey on multimodal medical signals fusion for smart healthcare systems. *Inf. Fusion*, 76:355–375.

Gonzales, T. I., Jeon, J. Y., Lindsay, T., Westgate, K., Perez-Pozuelo, I., Hollidge, S., Wijndaele, K., Rennie, K., Forouhi, N., Griffin, S., Wareham, N., and Brage, S. (2020a). Resting heart rate as a biomarker for tracking change in cardiorespiratory fitness of uk adults: The fenland study. *medRxiv*.

Gonzales, T. I., Westgate, K., Hollidge, S., Lindsay, T., Jeon, J., and Brage, S. (2020b). Estimating maximal oxygen consumption from heart rate response to submaximal ramped treadmill test. *medRxiv*.

Gonzales, T. I., Westgate, K., Strain, T., Hollidge, S., Jeon, J., Christensen, D. L., Jensen, J., Wareham, N. J., and Brage, S. (2021). Cardiorespiratory fitness assessment using risk-stratified exercise testing and dose-response relationships with disease outcomes. *Scientific Reports*, 11(1):15315.

Goswami, M., Szafer, K., Choudhry, A., Cai, Y., Li, S., and Dubrawski, A. (2024). Moment: a family of open time-series foundation models. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. (2012). A kernel two-sample test. 13(null):723–773.

Grill, J.-B., Strub, F., Altché, F., Tallec, C., Richemond, P. H., Buchatskaya, E., Doersch, C., Pires, B. A., Guo, Z. D., Azar, M. G., Piot, B., Kavukcuoglu, K., Munos, R., and Valko, M. (2020). Bootstrap your own latent a new approach to self-supervised learning. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Gui, J., Chen, T., Zhang, J., Cao, Q., Sun, Z., Luo, H., and Tao, D. (2024). A survey on self-supervised learning: Algorithms, applications, and future trends. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):9052–9071.

Guo, L. L., Pfohl, S. R., Fries, J., Johnson, A. E., Posada, J., Aftandilian, C., Shah, N., and Sung, L. (2022). Evaluation of domain generalization and adaptation on improving model robustness to temporal dataset shift in clinical medicine. *Scientific reports*, 12(1):2726.

Habib, F. A., Shakil, G. S., Iqbal, S. S. M., and Sajid, S. T. A. (2021). Self-diagnosis medical chatbot using artificial intelligence. In Goyal, D., Chaturvedi, P., Nagar, A. K.,

and Purohit, S., editors, *Proceedings of Second International Conference on Smart Energy and Communication*, pages 587–593, Singapore. Springer Singapore.

Hafner, M., Katsantoni, M., Köster, T., Marks, J., Mukherjee, J., Staiger, D., Ule, J., and Zavolan, M. (2021). Clip and complementary methods. *Nature Reviews Methods Primers*, 1(1):20.

Han, M., Qie, R., Shi, X., Yang, Y., Lu, J., Hu, F., Zhang, M., Zhang, Z., Hu, D., and Zhao, Y. (2022). Cardiorespiratory fitness and mortality from all causes, cardiovascular disease and cancer: dose–response meta-analysis of cohort studies. *British journal of sports medicine*, 56(13):733–739.

He, K., Fan, H., Wu, Y., Xie, S., and Girshick, R. (2020). Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Hemker, K., Simidjievski, N., and Jamnik, M. (2025). Healnet: multimodal fusion for heterogeneous biomedical data. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA. Curran Associates Inc.

Henriksen, A., Mikalsen, M., Woldaregay, A., Muzny, M., Hartvigsen, G., Hopstock, L., and Grimsgaard, S. (2018). Using fitness trackers and smartwatches to measure physical activity in research: Analysis of consumer wrist-worn wearables. *Journal of Medical Internet Research*, 20:e110.

Heydari, A. A., Gu, K., Srinivas, V., Yu, H., Zhang, Z., Zhang, Y., Paruchuri, A., He, Q., Palangi, H., Hammerquist, N., Metwally, A. A., Winslow, B., Kim, Y. H., Ayush, K., Yang, Y., Narayanswamy, G., Xu, M. A., Garrison, J., Lee, A. A., Vafeiadou, J., Graef, B., Galatzer-Levy, I. R., Schenck, E., Barakat, A., Perez, J., Shreibati, J., Hernandez, J., Faranesh, A. Z., Prieto, J. L., Heneghan, C., Liu, Y., Zhan, J., Malhotra, M., Patel, S. N., Althoff, T., Liu, X., McDuff, D., and XuhaiOrsonXu (2025). The anatomy of a personal health agent. *ArXiv*, abs/2508.20148.

Hill, N. R., Ayoubkhani, D., McEwan, P., Sugrue, D. M., Farooqui, U., Lister, S., Lumley, M., Bakhai, A., Cohen, A. T., O'Neill, M., et al. (2019). Predicting atrial fibrillation in primary care using machine learning. *PloS one*, 14(11):e0224582.

Hoffman, J., Tzeng, E., Park, T., Zhu, J.-Y., Isola, P., Saenko, K., Efros, A., and Darrell, T. (2018). CyCADA: Cycle-consistent adversarial domain adaptation. In Dy, J. and Krause, A., editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1989–1998. PMLR.

Huynh, T., Kornblith, S., Walter, M. R., Maire, M., and Khademi, M. (2022). Boosting contrastive self-supervised learning with false negative cancellation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2785–2795.

Iwase, S., Nakada, T., Shimada, T., Oami, T., Shimazui, T., Takahashi, N., Yamabe, J., Yamao, Y., and Kawakami, E. (2021). Prediction algorithm for icu mortality and length of stay using machine learning. *Scientific Reports*, 12.

Jahmunah, V., Oh, S. L., Wei, J. K. E., Ciaccio, E. J., Chua, K., San, T. R., and Acharya, U. R. (2019). Computer-aided diagnosis of congestive heart failure using ecg signals–a review. *Physica Medica*, 62:95–104.

Jaiswal, A., Babu, A. R., Zadeh, M. Z., Banerjee, D., and Makedon, F. (2020). A survey on contrastive self-supervised learning. *Technologies*, 9(1):2.

Jensen, K., Frydkjær, M., Jensen, N. M., Bannerholt, L. M., and Gam, S. (2021). A maximal rowing ergometer protocol to predict maximal oxygen uptake. *International Journal of Sports Physiology and Performance*, 16(3):382–386.

Ji, S., Zheng, X., and Wu, C. (2024). Hargpt: Are llms zero-shot human activity recognizers? *2024 IEEE International Workshop on Foundation Models for Cyber-Physical Systems & Internet of Things (FMSys)*, pages 38–43.

Jiang, W.-B., Zhao, L.-M., and Lu, B.-L. (2024). Large brain model for learning generic representations with tremendous eeg data in bci. *arXiv preprint arXiv:2405.18765*.

Jin, M., Wang, S., Ma, L., Chu, Z., Zhang, J. Y., Shi, X., Chen, P.-Y., Liang, Y., Li, Y.-F., Pan, S., and Wen, Q. (2024). Time-LLM: Time series forecasting by reprogramming large language models. In *The Twelfth International Conference on Learning Representations*.

Jo, Y.-Y., Lee, B. T., Kim, B. J., Hong, J., Lee, H. S., and myoung Kwon, J. (2024). New test-time scenario for biosignal: Concept and its approach. *ArXiv*, abs/2411.17785.

Johnson, A. E. W., Pollard, T. J., Shen, L., wei H. Lehman, L., Feng, M., Ghassemi, M. M., Moody, B., Szolovits, P., Celi, L. A., and Mark, R. G. (2016a). Mimic-iii, a freely accessible critical care database. *Scientific Data*, 3.

Johnson, A. E. W., Pollard, T. J., Shen, L., wei H. Lehman, L., Feng, M., Ghassemi, M. M., Moody, B., Szolovits, P., Celi, L. A., and Mark, R. G. (2016b). Mimic-iii, a freely accessible critical care database. *Scientific Data*, 3.

Kaptoge, S., Pennells, L., De Bacquer, D., Cooney, M. T., Kavousi, M., Stevens, G., Riley, L. M., Savin, S., Khan, T., Altay, S., Amouyel, P., Assmann, G., Bell, S., Ben-Shlomo,

Y., Berkman, L., Beulens, J. W., Björkelund, C., Blaha, M., Blazer, D. G., Bolton, T., Bonita Beaglehole, R., Brenner, H., Brunner, E. J., Casiglia, E., Chamnan, P., Choi, Y.-H., Chowdry, R., Coady, S., Crespo, C. J., Cushman, M., Dagenais, G. R., D'Agostino Sr, R. B., Daimon, M., Davidson, K. W., Engström, G., Ford, I., Gallacher, J., Gansevoort, R. T., Gaziano, T. A., Giampaoli, S., Grandits, G., Grimsgaard, S., Grobbee, D. E., Gudnason, V., Guo, Q., Tolonen, H., Humphries, S., Iso, H., Jukema, J. W., Kauhanen, J., Kengne, A. P., Khalili, D., Koenig, W., Kromhout, D., Krumholz, H., Lam, T. H., Laughlin, G., Marín Ibañez, A., Meade, T. W., Moons, K. G. M., Nietert, P. J., Ninomiya, T., Nordestgaard, B. G., O'Donnell, C., Palmieri, L., Patel, A., Perel, P., Price, J. F., Providencia, R., Ridker, P. M., Rodriguez, B., Rosengren, A., Roussel, R., Sakurai, M., Salomaa, V., Sato, S., Schöttker, B., Shara, N., Shaw, J. E., Shin, H.-C., Simons, L. A., Sofianopoulou, E., Sundström, J., Völzke, H., Wallace, R. B., Wareham, N. J., Willeit, P., Wood, D., Wood, A., Zhao, D., Woodward, M., Danaei, G., Roth, G., Mendis, S., Onuma, O., Varghese, C., Ezzati, M., Graham, I., Jackson, R., Danesh, J., and Di Angelantonio, E. (2019). World health organization cardiovascular disease risk charts: revised models to estimate risk in 21 global regions. *The Lancet Global Health*, 7(10):e1332–e1345.

Karimi, D., Dou, H., Warfield, S. K., and Gholipour, A. (2020). Deep learning with noisy labels: Exploring techniques and remedies in medical image analysis. *Medical Image Analysis*, 65:101759.

Kemp, B., Zwinderman, A. H., Tuk, B., Kamphuisen, H. A. C., and Oberye, J. J. L. (2000). Analysis of a sleep-dependent neuronal feedback loop: the slow-wave microcontinuity of the eeg. *IEEE Transactions on Biomedical Engineering*, 47:1185–1194.

Khan, Y., Ostfeld, A. E., Lochner, C. M., Pierre, A., and Arias, A. C. (2016). Monitoring of vital signs with flexible and wearable medical devices. *Advanced materials*, 28(22):4373–4395.

Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., and Krishnan, D. (2020). Supervised contrastive learning. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Kiyasseh, D., Zhu, T., and Clifton, D. A. (2020). Clocs: Contrastive learning of cardiac signals. In *International Conference on Machine Learning*.

Koroteev, M. V. (2021). Bert: a review of applications in natural language processing and understanding. *arXiv preprint arXiv:2103.11943*.

Kouw, W. M. (2018). An introduction to domain adaptation and transfer learning. *ArXiv*, abs/1812.11806.

Kwon, H., Wang, B., Abowd, G. D., and Plötz, T. (2021). Approaching the real-world: Supporting activity recognition training with virtual imu data. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 5(3):1–32.

Lan, X., Ng, D., linda Qiao, and Feng, M. (2021). Intra-inter subject self-supervised learning for multivariate cardiac signals. In *AAAI Conference on Artificial Intelligence*.

Laukkanen, J. A., Lakka, T. A., Rauramaa, R., Kuhanen, R., Venäläinen, J. M., Salonen, R., and Salonen, J. T. (2001). Cardiovascular Fitness as a Predictor of Mortality in Men. *Archives of Internal Medicine*, 161(6):825–831.

Lee, H.-C., Park, Y., Yoon, S. B., Yang, S. M., Park, D., and Jung, C.-W. (2022). Vitaldb, a high-fidelity multi-parameter vital signs database in surgical patients. *Scientific Data*, 9(1):279.

Liang, Y., Wen, H., Nie, Y., Jiang, Y., Jin, M., Song, D., Pan, S., and Wen, Q. (2024). Foundation models for time series analysis: A tutorial and survey. *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*.

Lin, W., Sun, M.-T., Poovandran, R., and Zhang, Z. (2008). Human activity recognition for video surveillance. In *2008 IEEE international symposium on circuits and systems (ISCAS)*, pages 2737–2740. IEEE.

Lindsay, T., Westgate, K., Wijndaele, K., Hollidge, S., Kerrison, N., Forouhi, N., Griffin, S., Wareham, N., and Brage, S. (2019a). Descriptive epidemiology of physical activity energy expenditure in uk adults (the fenland study). *International Journal of Behavioral Nutrition and Physical Activity*, 16.

Lindsay, T., Westgate, K., Wijndaele, K., Hollidge, S., Kerrison, N., Forouhi, N., Griffin, S., Wareham, N., and Brage, S. (2019b). Descriptive epidemiology of physical activity energy expenditure in uk adults (the fenland study). *International Journal of Behavioral Nutrition and Physical Activity*, 16.

Lipton, Z. C., Wang, Y., and Smola, A. J. (2018). Detecting and correcting for label shift with black box predictors. *CoRR*, abs/1802.03916.

Liu, H., Cui, S., Zhao, X., and Cong, F. (2023). Detection of obstructive sleep apnea from single-channel ecg signals using a cnn-transformer architecture. *Biomedical Signal Processing and Control*, 82:104581.

Liu, H., Wang, J., and Long, M. (2021). Cycle self-training for domain adaptation. *CoRR*, abs/2103.03571.

Liu, P., Guo, H., Dai, T., Li, N., Bao, J., Ren, X., Jiang, Y., and Xia, S.-T. (2025). Calf: aligning llms for time series forecasting via cross-modal fine-tuning. In *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence and Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence and Fifteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'25/IAAI'25/EAAI'25. AAAI Press.

Liu, Y., Butkow, K.-J., Stuchbury-Wass, J., Pullin, A., Ma, D., and Mascolo, C. (2024). Respear: Earable-based robust respiratory rate monitoring. *2025 IEEE International Conference on Pervasive Computing and Communications (PerCom)*, pages 67–77.

Long, M., Cao, Y., Wang, J., and Jordan, M. (2015). Learning transferable features with deep adaptation networks. In Bach, F. and Blei, D., editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 97–105, Lille, France. PMLR.

Loni, M., Poursalim, F., Asadi, M., and Gharehbaghi, A. (2024). A review on generative ai models for synthetic medical text, time series, and longitudinal data. *NPJ Digital Medicine*, 8.

Luo, D., Cheng, W., Wang, Y., Xu, D., Ni, J., Yu, W., Zhang, X., Liu, Y., Chen, Y., Chen, H., and Zhang, X. (2023). Time series contrastive learning with information-aware augmentations. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'23/IAAI'23/EAAI'23. AAAI Press.

Mallick, S. and Baths, V. (2024). Novel deep learning framework for detection of epileptic seizures using eeg signals. *Frontiers in Computational Neuroscience*, 18.

Mandsager, K., Harb, S., Cremer, P., Phelan, D., Nissen, S. E., and Jaber, W. (2018). Association of cardiorespiratory fitness with long-term mortality among adults undergoing exercise treadmill testing. *JAMA network open*, 1(6):e183605–e183605.

Marlin, B. M., Kale, D. C., Khemani, R. G., and Wetzel, R. C. (2012). Unsupervised pattern discovery in electronic health care data using probabilistic clustering models. In *Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium*, IHI '12, page 389–398, New York, NY, USA. Association for Computing Machinery.

Mathelin, A., Richard, G., Mougeot, M., and Vayatis, N. (2020). Adversarial weighting for domain adaptation in regression. *2021 IEEE 33rd International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 49–56.

Mathew, G., Barbosa, D., Prince, J., and Venkatraman, S. (2024). Foundation models for cardiovascular disease detection via biosignals from digital stethoscopes. *npj Cardiovascular Health*, 1(1):25.

McKeen, K., Masood, S., Toma, A., Rubin, B., and Wang, B. (2024). Ecg-fm: An open electrocardiogram foundation model. *arXiv preprint arXiv:2408.05178*.

Mei, K., Zhu, C., Jiang, L., Liu, J., and Qiao, Y. (2020). Cross-stained segmentation from renal biopsy images using multi-level adversarial learning. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1424–1428.

Meng, Q., Qian, H., Liu, Y., Cui, L., Xu, Y., and Shen, Z. (2023). Mhccl: masked hierarchical cluster-wise contrastive learning for multivariate time series. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'23/IAAI'23/EAAI'23. AAAI Press.

Mohsenvand, M. N., Izadi, M. R., and Maes, P. (2020). Contrastive representation learning for electroencephalogram classification. In Alsentzer, E., McDermott, M. B. A., Falck, F., Sarkar, S. K., Roy, S., and Hyland, S. L., editors, *Proceedings of the Machine Learning for Health NeurIPS Workshop*, volume 136 of *Proceedings of Machine Learning Research*, pages 238–253. PMLR.

Moody, G. (1983). A new method for detecting atrial fibrillation using rr intervals. *Proc. Comput. Cardiol.*, 10:227–230.

Morokuma, S., Hayashi, T., Kanegae, M., Mizukami, Y., Asano, S., Kimura, I., Tateizumi, Y., Ueno, H., Ikeda, S., and Niizeki, K. (2023). Deep learning-based sleep stage classification with cardiorespiratory and body movement activities in individuals with suspected sleep disorders. *Scientific Reports*, 13.

Mushtaq, R. (2011). Augmented dickey fuller test. *Econometrics: Mathematical Methods & Programming eJournal*.

Nakamura, T., Kiyono, K., Wendt, H., Abry, P., and Yamamoto, Y. (2016). Multiscale analysis of intensive longitudinal biomedical signals and its clinical applications. *Proceedings of the IEEE*, 104(2):242–261.

Narayanswamy, G., Liu, X., Ayush, K., Yang, Y., Xu, X., Liao, S., Garrison, J., Tailor, S., Sunshine, J., Liu, Y., Althoff, T., Narayanan, S., Kohli, P., Zhan, J., Malhotra, M., Patel, S. N., Abdel-Ghaffar, S., and McDuff, D. (2024). Scaling wearable foundation models. *ArXiv*, abs/2410.13638.

Nelson, B. W., Low, C. A., Jacobson, N. C., Areán, P. A., Torous, J. B., and Allen, N. B. (2020). Guidelines for wrist-worn consumer wearable assessment of heart rate in biobehavioral research. *NPJ Digital Medicine*, 3.

Nes, B., Janszky, I., Vatten, L., Nilsen, T., Aspenes, S., and Wisloff, U. (2011). Estimating v˙o2peak from a nonexercise prediction model: The hunt study, norway. *Medicine and science in sports and exercise*, 43:2024–30.

Niu, L., Chen, C., Liu, H., Zhou, S., and Shu, M. (2020). A deep-learning approach to ecg classification based on adversarial domain adaptation. In *Healthcare*, volume 8, page 437. MDPI.

Pan, Z., Jiang, Y., Garg, S., Schneider, A., Nevmyvaka, Y., and Song, D. (2024). S2ip-llm: semantic space informed prompt learning with llm for time series forecasting. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Park, S., Lee, J., Lee, P., Hwang, S., Kim, D., and Byun, H. (2022). Fair Contrastive Learning for Facial Attribute Classification . In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10379–10388, Los Alamitos, CA, USA. IEEE Computer Society.

Perez-Pozuelo, I., Spathis, D., Clifton, E., and Mascolo, C. (2021). *Wearables, smartphones, and artificial intelligence for digital phenotyping and health*, pages 33–54.

Perone, C. S., Ballester, P., Barros, R. C., and Cohen-Adad, J. (2019). Unsupervised domain adaptation for medical imaging segmentation with self-ensembling. *NeuroImage*, 194:1–11.

Plasqui, G. and Westerterp, K. (2006). Accelerometry and heart rate as a measure of physical fitness. *Medicine and science in sports and exercise*, 38:1510–4.

Qayyum, A., Qadir, J., Bilal, M., and Al-Fuqaha, A. (2020). Secure and robust machine learning for healthcare: A survey. *IEEE Reviews in Biomedical Engineering*, 14:156–180.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. (2019). Language models are unsupervised multitask learners. Technical report, OpenAI.

Rajkomar, A., Oren, E., Chen, K., Dai, A. M., Hajaj, N., Hardt, M., Liu, P. J., Liu, X., Marcus, J., Sun, M., Sundberg, P., Yee, H., Zhang, K., Zhang, Y., Flores, G., Duggan, G. E., Irvine, J., Le, Q. V., Litsch, K., Mossin, A., Tansuwan, J., Wang, D., Wexler, J., Wilson, J., Ludwig, D., Volchenboum, S. L., Chou, K., Pearson, M., Madabushi, S., Shah, N. H., Butte, A. J., Howell, M. D., Cui, C., Corrado, G. S., and Dean, J. (2018). Scalable and accurate deep learning with electronic health records. *NPJ Digital Medicine*, 1.

Raschka, S. (2018). Model evaluation, model selection, and algorithm selection in machine learning. *CoRR*, abs/1811.12808.

Rietjens, G., Kuipers, H., Kester, A., and Keizer, H. (2001). Validation of a computerized metabolic measurement system during low and high intensity exercise1. *International journal of sports medicine*, 22:291–4.

Robinson, J. D., Chuang, C.-Y., Sra, S., and Jegelka, S. (2021). Contrastive learning with hard negative samples. In *International Conference on Learning Representations*.

Rubanova, Y., Chen, R. T. Q., and Duvenaud, D. (2019). *Latent ODEs for irregularly-sampled time series*. Curran Associates Inc., Red Hook, NY, USA.

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, 323:533–536.

Sabry, F., Eltaras, T. A., Labda, W., Alzoubi, K., and Malluhi, Q. M. (2022). Machine learning for healthcare wearable devices: The big picture. *Journal of Healthcare Engineering*, 2022.

Saeed, A., Grangier, D., and Zeghidour, N. (2020). Contrastive learning of general-purpose audio representations. *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3875–3879.

Saeed, A., Ozcelebi, T., and Lukkien, J. (2019). Multi-task self-supervised learning for human activity detection. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 3(2):1–30.

Saito, K., Kim, D., Sclaroff, S., Darrell, T., and Saenko, K. (2019). Semi-supervised domain adaptation via minimax entropy. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8050–8058.

Sanabria, A. R., Zambonelli, F., Dobson, S., and Ye, J. (2021). Contrasgan: Unsupervised domain adaptation in human activity recognition via adversarial and contrastive learning. *Pervasive and Mobile Computing*, 78:101477.

Schirmer, M., Eltayeb, M., Lessmann, S., and Rudolph, M. (2022). Modeling irregular time series with continuous recurrent units. In Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., and Sabato, S., editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 19388–19405. PMLR.

Schultz, T. and Maedche, A. (2023). Biosignals meet adaptive systems. *SN Applied Sciences*, 5.

Shcherbina, A., Mattsson, C., Waggott, D., Salisbury, H., Christle, J., Hastie, T., Wheeler, M., and Ashley, E. (2017). Accuracy in wrist-worn, sensor-based measurements of heart rate and energy expenditure in a diverse cohort. *Journal of Personalized Medicine*, 7:3.

Shen, J., Qu, Y., Zhang, W., and Yu, Y. (2018). Wasserstein distance guided representation learning for domain adaptation. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'18/IAAI'18/EAAI'18. AAAI Press.

Sherstinsky, A. (2020). Fundamentals of recurrent neural network (rnn) and long short-term memory (lstm) network. *Physica D: Nonlinear Phenomena*, 404:132306.

Shi, P., Ye, W., and Qin, Z. (2021). Self-supervised pre-training for time series classification. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8.

Shoeibi, A., Khodatars, M., Ghassemi, N., Jafari, M., Moridian, P., Alizadehsani, R., Panahiazar, M., Khozeimeh, F., Zare, A., Hosseini-Nejad, H., Khosravi, A., Atiya, A. F., Aminshahidi, D., Hussain, S., Rouhani, M., Nahavandi, S., and Acharya, U. R. (2020). Epileptic seizures detection using deep learning techniques: A review. *International Journal of Environmental Research and Public Health*, 18.

Shukla, S. N. and Marlin, B. (2021). Multi-time attention networks for irregularly sampled time series. In *International Conference on Learning Representations*.

Shukla, S. N. and Marlin, B. M. (2020). A survey on principles, models and methods for learning from irregularly sampled time series: From discretization to attention and invariance. *ArXiv*, abs/2012.00168.

Siebra, C. A., Kurpicz-Briki, M., and Wac, K. (2024). Transformers in health: a systematic review on architectures for longitudinal data analysis. *Artif. Intell. Rev.*, 57:32.

Silva, I., Moody, G. B., Scott, D. J., Celi, L. A., and Mark, R. G. (2012). Predicting in-hospital mortality of icu patients: The physionet/computing in cardiology challenge 2012. *2012 Computing in Cardiology*, pages 245–248.

Singhal, K., Azizi, S., Tu, T., Mahdavi, S., Wei, J., Chung, H. W., Scales, N., Tanwani, A. K., Cole-Lewis, H. J., Pfohl, S. J., Payne, P. A., Seneviratne, M. G., Gamble, P., Kelly, C., Scharli, N., Chowdhery, A., Mansfield, P. A., y Arcas, B. A., Webster, D. R., Corrado, G. S., Matias, Y., Chou, K. H.-L., Gottweis, J., Tomaev, N., Liu, Y., Rajkomar, A., Barral, J. K., Semturs, C., Karthikesalingam, A., and Natarajan, V. (2022). Large language models encode clinical knowledge. *Nature*, 620:172 – 180.

Spathis, D., Perez-Pozuelo, I., Brage, S., Wareham, N. J., and Mascolo, C. (2021). Self-supervised transfer learning of physiological representations from free-living wearable data. In *Proceedings of the Conference on Health, Inference, and Learning*. ACM.

Spathis, D., Perez-Pozuelo, I., Gonzales, T. I., Wu, Y., Brage, S., Wareham, N. J., and Mascolo, C. (2022). Longitudinal cardio-respiratory fitness prediction through wearables in free-living environments. *NPJ Digital Medicine*, 5.

Srivastava, N., Mansimov, E., and Salakhutdinov, R. (2015). Unsupervised learning of video representations using lstms. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, page 843–852. JMLR.org.

Stegle, O., Fallert, S. V., MacKay, D. J. C., and Brage, S. (2008). Gaussian process robust regression for noisy heart rate data. *IEEE Transactions on Biomedical Engineering*, 55(9):2143–2151.

Sun, B. and Saenko, K. (2016). Deep coral: Correlation alignment for deep domain adaptation. In *European conference on computer vision*, pages 443–450. Springer.

Sun, W., Zhang, J., Wang, J., Liu, Z., Zhong, Y., Feng, T., Guo, Y., Zhang, Y., and Barnes, N. (2023). Learning audio-visual source localization via false negative aware contrastive learning. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6420–6429.

Swain, D. P., Brawner, C. A., of Sports Medicine, A. C., et al. (2014). *ACSM's resource manual for guidelines for exercise testing and prescription*. Wolters Kluwer Health/Lippincott Williams & Wilkins.

Tang, C. I., Perez-Pozuelo, I., Spathis, D., Brage, S., Wareham, N., and Mascolo, C. (2021). Selfhar: Improving human activity recognition through self-training with unlabeled data. *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies*, 5(1):1–30.

Tay, Y., Bahri, D., Yang, L., Metzler, D., and Juan, D.-C. (2020). Sparse sinkhorn attention. In *International conference on machine learning*, pages 9438–9447. PMLR.

Tipirneni, S. and Reddy, C. K. (2022). Self-supervised transformer for sparse and irregularly sampled multivariate clinical time-series. *ACM Trans. Knowl. Discov. Data*, 16(6).

Tonekaboni, S., Eytan, D., and Goldenberg, A. (2021). Unsupervised representation learning for time series with temporal neighborhood coding. *arXiv preprint arXiv:2106.00750*.

Turmon, M. J. and Fine, T. L. (1994). Sample size requirements for feedforward neural networks. In *Neural Information Processing Systems*.

Tzeng, E., Hoffman, J., Saenko, K., and Darrell, T. (2017). Adversarial discriminative domain adaptation. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2962–2971.

Tzeng, E., Hoffman, J., Zhang, N., Saenko, K., and Darrell, T. (2014). Deep domain confusion: Maximizing for domain invariance. *ArXiv*, abs/1412.3474.

UK Biobank (2011). Uk biobank cardio assessment manual version 1.0. `https://biobank.ctsu.ox.ac.uk/crystal/crystal/docs/Cardio.pdf`.

Uth, N., Sørensen, H., Overgaard, K., and Pedersen, P. K. (2004). Estimation of vo2max from the ratio between hrmax and hrrest–the heart rate ratio method. *European journal of applied physiology*, 91(1):111–115.

van den Oord, A., Li, Y., and Vinyals, O. (2018). Representation learning with contrastive predictive coding. *ArXiv*, abs/1807.03748.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.

Venkataramani, R., Ravishankar, H., and Anamandra, S. (2018). Towards continuous domain adaptation for healthcare. *ArXiv*, abs/1812.01281.

Verma, V., Luong, T., Kawaguchi, K., Pham, H., and Le, Q. (2021). Towards domain-agnostic contrastive learning. In Meila, M. and Zhang, T., editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 10530–10541. PMLR.

Vidulin, Vedrana, L. M. K. B. P. R. and Krivec, J. (2010). Localization Data for Person Activity. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C57G8X.

Vrigkas, M., Nikou, C., and Kakadiaris, I. A. (2015). A review of human activity recognition methods. *Frontiers in Robotics and AI*, 2:28.

Wan, Z., Li, M., Liu, S., Huang, J., Tan, H., and Duan, W. (2023). Eegformer: A transformer–based brain activity classification method using eeg signal. *Frontiers in neuroscience*, 17:1148855.

Wang, D., Shelhamer, E., Liu, S., Olshausen, B. A., and Darrell, T. (2021). Tent: Fully test-time adaptation by entropy minimization. In *International Conference on Learning Representations*.

Wang, M. and Deng, W. (2018). Deep visual domain adaptation: A survey. *CoRR*, abs/1802.03601.

Wang, S., Wang, J., Xi, H., Zhang, B., Zhang, L., and Wei, H. (2024a). Optimization-free test-time adaptation for cross-person activity recognition. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 7(4):1–27.

Wang, W., Dai, J., Chen, Z., Huang, Z., Li, Z., Zhu, X., Hu, X., Lu, T., Lu, L., Li, H., et al. (2023a). Internimage: Exploring large-scale vision foundation models with deformable convolutions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14408–14419.

Wang, Y., Han, Y., Wang, H., and Zhang, X. (2023b). Contrast everything: a hierarchical contrastive framework for medical time-series. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Wang, Y., Wu, H., Dong, J., Liu, Y., Long, M., and Wang, J. (2024b). Deep time series models: A comprehensive survey and benchmark. *ArXiv*, abs/2407.13278.

Weerakody, P. B., Wong, K. W., Wang, G., and Ela, W. (2021). A review of irregular time series data handling with gated recurrent neural networks. *Neurocomputing*, 441:161–178.

Wen, Q., Zhou, T., Zhang, C., Chen, W., Ma, Z., Yan, J., and Sun, L. (2022). Transformers in time series: A survey. *arXiv preprint arXiv:2202.07125*.

White, T., Westgate, K., Wareham, N., and Brage, S. (2016). Estimation of physical activity energy expenditure during free-living from wrist accelerometry in uk adults. *PLOS ONE*, 11:e0167472.

World Health Organization (2022). Ageing and health. `https://www.who.int/news-room/fact-sheets/detail/ageing-and-health`. Accessed: 2025-07-23.

Wu, H., Xu, J., Wang, J., and Long, M. (2021). Autoformer: decomposition transformers with auto-correlation for long-term series forecasting. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, NIPS '21, Red Hook, NY, USA. Curran Associates Inc.

Wu, Y., Dang, T., Spathis, D., Jia, H., and Mascolo, C. (2024). Statiocl: Contrastive learning for time series via non-stationary and temporal contrast. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, CIKM '24, page 2575–2584, New York, NY, USA. Association for Computing Machinery.

Wu, Y., Spathis, D., Jia, H., Perez-Pozuelo, I., Gonzales, T. I., Brage, S., Wareham, N., and Mascolo, C. (2023). Udama: Unsupervised domain adaptation through multi-discriminator adversarial training with noisy labels improves cardio-fitness prediction. In Deshpande, K., Fiterau, M., Joshi, S., Lipton, Z., Ranganath, R., Urteaga, I., and Yeung, S., editors, *Proceedings of the 8th Machine Learning for Healthcare Conference*, volume 219 of *Proceedings of Machine Learning Research*, pages 863–883. PMLR.

Wu, Y., Spathis, D., Jia, H., Perez-Pozuelo, I., Gonzales, T. I., Brage, S., Wareham, N. J., and Mascolo, C. (2022). Turning silver into gold: Domain adaptation with noisy labels for wearable cardio-respiratory fitness prediction. *ArXiv*, abs/2211.10475.

Yalavarthi, V. K., Madhusudhanan, K., Scholz, R., Ahmed, N., Burchert, J., Jawed, S., Born, S., and Schmidt-Thieme, L. (2024). Grafiti: graphs for forecasting irregularly sampled time series. In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'24/IAAI'24/EAAI'24. AAAI Press.

Yan, H., Ding, Y., Li, P., Wang, Q., Xu, Y., and Zuo, W. (2017). Mind the class weight bias: Weighted maximum mean discrepancy for unsupervised domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2272–2281.

Yang, J., Nguyen, M. N., San, P. P., Li, X., and Krishnaswamy, S. (2015). Deep convolutional neural networks on multichannel time series for human activity recognition. In *International Joint Conference on Artificial Intelligence*.

Yang, K., Hong, M., Zhang, J., Luo, Y., Su, Y., Zhang, O., Yu, X., Zhou, J., Yang, L., Qian, M., Zhang, P., and Nie, Z. (2024). Ecg-lm: Understanding electrocardiogram with a large language model. *Health Data Science*, 5.

Yang, X., He, X., Liang, Y., Yang, Y., Zhang, S., and Xie, P. (2020). Transfer learning or self-supervised learning? a tale of two pretraining paradigms. *ArXiv*, abs/2007.04234.

Yang, Y., Liu, X., Wu, J., Borac, S., Katabi, D., Poh, M., and McDuff, D. (2023). Simper: Simple self-supervised learning of periodic targets. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Ye, X., Sakurai, K., Nair, N.-K. C., and Wang, K. I.-K. (2024). Machine learning techniques for sensor-based human activity recognition with data heterogeneity—a review. *Sensors*, 24(24):7975.

Ying, X. (2019). An overview of overfitting and its solutions. In *Journal of physics: Conference series*, volume 1168, page 022022. IOP Publishing.

Yu, H., Guo, P., and Sano, A. (2024). Ecg semantic integrator (esi): A foundation ecg model pretrained with llm-enhanced cardiological text. *ArXiv*, abs/2405.19366.

Yu, Y., Si, X., Hu, C., and Zhang, J. (2019). A review of recurrent neural networks: Lstm cells and network architectures. *Neural computation*, 31(7):1235–1270.

Yuan, H., Chan, S., Creagh, A. P., Tong, C., Clifton, D. A., and Doherty, A. (2022). Self-supervised learning for human activity recognition using 700,000 person-days of wearable data. *NPJ Digital Medicine*, 7.

Yue, Z., Wang, Y., Duan, J., Yang, T., Huang, C., Tong, Y., and Xu, B. (2021). Ts2vec: Towards universal representation of time series. In *AAAI Conference on Artificial Intelligence*.

Zadorozhny, K., Thoral, P., Elbers, P., and Cinà, G. (2022). Out-of-distribution detection for medical applications: Guidelines for practical evaluation. In *Multimodal AI in healthcare: A paradigm shift in health intelligence*, pages 137–153. Springer.

Zhang, D., Yuan, Z., Yang, Y., Chen, J., Wang, J., and Li, Y. (2023a). Brant: Foundation model for intracranial neural signal. In *Neural Information Processing Systems*.

Zhang, H., Li, F., Liu, S., Zhang, L., Su, H., Zhu, J., Ni, L. M., and Shum, H.-Y. (2022a). Dino: Detr with improved denoising anchor boxes for end-to-end object detection. *arXiv preprint arXiv:2203.03605*.

Zhang, J., Zheng, S., Cao, W., Bian, J., and Li, J. (2023b). Warpformer: A multi-scale modeling approach for irregular clinical time series. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '23, page 3273–3285, New York, NY, USA. Association for Computing Machinery.

Zhang, S., Li, Y., Zhang, S., Shahabi, F., Xia, S., Deng, Y., and Alshurafa, N. (2022b). Deep learning in human activity recognition with wearable sensors: A review on advances. *Sensors*, 22(4):1476.

Zhang, W., Yin, C., Liu, H., Zhou, X., and Xiong, H. (2024a). Irregular multivariate time series forecasting: A transformable patching graph neural networks approach. In *Forty-first International Conference on Machine Learning*.

Zhang, X., Chowdhury, R. R., Gupta, R. K., and Shang, J. (2024b). Large language models for time series: A survey. *ArXiv*, abs/2402.01801.

Zhang, X., Zeman, M., Tsiligkaridis, T., and Zitnik, M. (2022c). Graph-guided network for irregularly sampled multivariate time series. In *International Conference on Learning Representations*.

Zhang, X., Zhao, Z., Tsiligkaridis, T., and Zitnik, M. (2022d). Self-supervised contrastive pre-training for time series via time-frequency consistency. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Zhang, Y., Yang, X., Ivy, J., and Chi, M. (2019). Attain: Attention-based time-aware lstm networks for disease progression modeling. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 4369–4375. International Joint Conferences on Artificial Intelligence Organization.

Zhao, H., Zhang, S., Wu, G., Moura, J. M. F., Costeira, J. P., and Gordon, G. J. (2018). Adversarial multiple source domain adaptation. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Zhao, Z., Alzubaidi, L., Zhang, J., Duan, Y., and Gu, Y. (2024). A comparison review of transfer learning and self-supervised learning: Definitions, applications, advantages and limitations. *Expert Systems with Applications*, 242:122807.

Zhou, T., Niu, P., Wang, X., Sun, L., and Jin, R. (2023). One fits all: Power general time series analysis by pretrained LM. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Appendix A

Basics of deep learning

A.1 Deep neural networks: fundamentals and training pipeline

The core aim of a neural network is to approximate a function $f(\cdot)$ that maps an input $\mathbf{x}$ to a target output $\mathbf{y}$:

$$\hat{\mathbf{y}} = f(\mathbf{x}; \boldsymbol{\theta}), \tag{A.1}$$

where $\boldsymbol{\theta}$ denotes the set of trainable parameters (weights and biases) of the network. A deep neural network (DNN) consists of multiple layers of nonlinear transformations, stacked to progressively extract higher-level features from the input. Formally, for a network composed of L layers:

$$f(\mathbf{x}) = f^{(L)}\left(f^{(L-1)}\left(\dots f^{(1)}(\mathbf{x})\right)\right), \tag{A.2}$$

where $f^{(l)}(\cdot)$ denotes the operation of the l-th layer.

Each layer applies an affine transformation followed by a nonlinear activation function:

$$\mathbf{z}^{(l)} = g\left(\mathbf{W}^{(l)}\mathbf{h}^{(l-1)} + \mathbf{b}^{(l)}\right), \tag{A.3}$$

$$\mathbf{h}^{(0)} = \mathbf{x}, \tag{A.4}$$

where $\mathbf{W}^{(l)}$ and $\mathbf{b}^{(l)}$ are the weight matrix and bias vector of the l-th layer, $g(\cdot)$ is a nonlinear activation such as the Rectified Linear Unit (ReLU), and $\mathbf{h}^{(l)}$ is the output of the l-th layer.

Training via backpropagation. The network parameters $\boldsymbol{\theta}$ are learned by minimizing a task-specific loss function $\mathcal{L}$, such as mean squared error for regression or cross-entropy

for classification, using a labelled dataset $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^N$. The loss over the dataset is:

$$\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{N} \sum_{i=1}^{N} \ell\left(f(\mathbf{x}_i; \boldsymbol{\theta}), \mathbf{y}_i\right), \tag{A.5}$$

where $\ell(\cdot, \cdot)$ denotes the pointwise loss.

The optimisation is performed using gradient-based methods, where gradients of $\mathcal{L}$ with respect to $\boldsymbol{\theta}$ are computed via the backpropagation algorithm:

$$\frac{\partial \mathcal{L}}{\partial \mathbf{W}^{(l)}} = \frac{\partial \mathcal{L}}{\partial \mathbf{h}^{(l)}} \cdot \frac{\partial \mathbf{h}^{(l)}}{\partial \mathbf{W}^{(l)}}. \tag{A.6}$$

The parameters are then updated iteratively:

$$\mathbf{W}^{(l)} \leftarrow \mathbf{W}^{(l)} - \eta \frac{\partial \mathcal{L}}{\partial \mathbf{W}^{(l)}}, \tag{A.7}$$

where η is the learning rate.