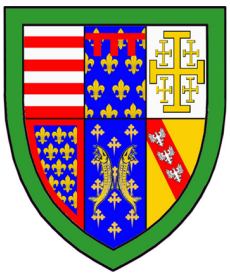
UNIVERSITY OF CAMBRIDGE
Department of Computer Science and Technology

Reliable and decentralised deep learning for physiological data

Tong Xia



Queens' College

January 2024

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 specified in the text. It is not substantially the same as any that I have submitted, or, is being concurrently submitted for a degree or 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. I further state that no substantial part of my thesis has already been submitted, or, is being concurrently submitted for any such 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 of 60 000 for the relevant Degree Committee.

Tong Xia

April 2024

Abstract

Reliable and decentralised deep learning for physiological data

Tong Xia

Physiological data encompass measurements from various bodily functions and processes. By employing machine learning to model these data, especially with the advancement of mobile sensing technologies, it becomes feasible to automatically and continually monitor and diagnose one's health status. This holds considerable promise for easing the burden on clinical resources and ensuring timely treatment for the wider population. Nonetheless, significant challenges related to the data and the modelling methods are yet to be resolved, obstructing the deployment of machine learning, especially deep learning, in real-world healthcare contexts.

One challenge is that labelled physiological data for model development are usually insufficient and imbalanced, leading to models occasionally exhibiting bias and overconfidence in their predictions. This can result in unreliable diagnoses which yield expensive clinical costs. Moreover, deep learning research generally requires massive data on a centralised server, while privacy concerns hinder the aggregation of physiological data from individuals or hospitals.

In order to tackle these challenges and pave the way for reliable deep learning-driven health diagnostics, this thesis proposes several novel solutions and makes the following contributions:

Chapter 4 introduces an ensemble learning approach designed to handle data imbalance and model overconfidence for *binary health screening*. This method utilises balanced training sets derived from imbalanced physiological data, training multiple ensemble models. The predictions from these models are fused to reduce bias and calibrate confidence from a signal model, with model uncertainty measured by the inconsistency among multiple models. This approach effectively mitigates model overconfidence, thereby facilitating reliable automated diagnoses.

In Chapter 5, an efficient uncertainty quantification approach is presented to improve the reliability of *multi-class mobile health diagnostics*. This approach incorporates the cutting-edge

technique of evidential deep learning and introduces two novel mechanisms specifically designed to handle class imbalance. The quantified uncertainty enables accurate and efficient detection of misdiagnoses and out-of-training distributed inputs.

Chapter 6 introduces a *cross-device* federated learning method to address privacy concerns arising from gathering physiological data for model development. This method allows physiological data to remain on personal mobile devices, with only locally trained models aggregated into a global health diagnostic model. To mitigate bias caused by data imbalance, a novel loss-weighted model aggregation method is proposed to enhance the performance of the global model.

Chapter 7 illustrates a *cross-silo* federated learning method that enables multiple data holders such as hospitals to collaboratively train a model without exchanging raw data. The distributional heterogeneity of these physiological data silos poses a challenge to federated learning. To address this, a novel method based on feature sharing and augmentation is proposed to balance privacy protection and model performance.

All proposed methods have been validated using real-world physiological datasets and commonly used machine learning benchmark data. Specific attention is given to clinical tasks, including the modelling of respiratory audio for respiratory health screening, ECG signals for predicting cardiovascular diseases, and dermoscopic images for detecting skin cancer. Extensive experiments demonstrate that these methods effectively address challenges posed by limited, imbalanced, and decentralised physiological data, thereby enabling reliable health diagnoses. These contributions have significant potential to advance the deployment of deep learning in real-world healthcare scenarios.

Acknowledgements

During my Ph.D. study, if there is one person to thank, I will definitely and firstly convey my most heartfelt gratitude to my supervisor, Prof. Cecilia Mascolo. Researchers treasure the University of Cambridge as the nurturing soil for them to grow, then my supervisor is the daily rising sun, warming and guiding my way all the time. Without her, I cannot be currently what I am. Her high academic standards, effective time management skills, and love and care for students, these wonderful qualities will be my everlasting goals in my life, spanning beyond my four-year doctoral journey.

I would also like to extend my gratitude to everyone in the Mobile Systems Research Laboratory for creating a welcoming and friendly atmosphere. Their collective efforts have contributed to a stimulating and collaborative research environment. Especially, I want to express my appreciation to Dr. Jing Han, Dr. Ting Dang, and Dr. Abhirup Ghosh, for their unwavering support throughout my study. Their academic guidance has been invaluable in shaping my research and personal growth. Additionally, I am deeply thankful for the thought-provoking discussions I had with Dr. Jagmohan Chauhan, Dr. Dimitris Spathis, and Dr. Lorena Qendro. Moreover, I would like to offer special thanks to Dr. Yang Liu, Dr. Hong Jia, Dr. Qiang Yang, Yu Wu, and Yuwei Zhang, as well as my fellow Young Kwon and Kayla-Jade Butkow for their emotional support and companionship during the challenging moments of my Ph.D. study. Their presence has provided me with the strength and encouragement needed to overcome every dark moment in these four years.

My sincere thanks to other faculties at the Department of Computer Science and Technology including Prof. Pietro Lio, Prof. Nicholas D. Lane and Dr. Carl Henrik Ek, for their great guidance and help during the progress of my Ph.D. research. In the final stage of my Ph.D. journey, my external examiner, Prof. David Clifton, from the University of Oxford, has provided me with numerous pieces of constructive feedback. This not only enabled me to produce a higher quality doctoral thesis but also inspired me to set a high standard for my future research.

The college culture in Cambridge is truly unique, and my three course years at Queens' College have left me with unforgettable memories. I am deeply thankful for the support of my college tutor, Prof. Andrew Marsham. Living in college accommodations with my college mates, Di

Zhao and Julie Matte, was a wonderful experience, and I would also like to express my gratitude to them.

I feel so lucky to have friends from Cambridge and the UK Tsinghua Alumni Association (UKTA). I have greatly enjoyed playing badminton and participating in various activities with Gao Yang, Siyuan Guo and Xinchi Qiu in West Cambridge. Our bond extends beyond the cherished experiences we shared in the past, and continues to grow through a lasting and ever-renewing friendship. Besides, I express my gratitude to the UKTA and its committee members. Through their organisation of numerous meaningful and joyful activities, they have enriched my spare time. As a committee member, I have also experienced the joy of assisting others and have had the pleasure of meeting many new alumni who are kind and supportive. Through these activities, I have built strong friendships with many alumni in the UK, such as Wenyu Wang, Qiumeng Li, and Fengning Yang.

I also want to give my deep thanks to my parents for their constant support and understanding. My Ph.D. started when the COVID-19 pandemic spread all over the world. Although with great concern for me, the only child in my family, my parents always respect my choice and give me full emotional support, allowing me to pursue my dream. I love them forever.

As a closure, I voice my appreciation to myself. From a Chinese girl born in a small town, who cannot speak English fluently, to a Ph.D. student at the University of Cambridge, majoring in computer science, I like to challenge myself, and I will continue challenging myself to be a greater researcher and person.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

AI will revolutionise the way we live, including our healthcare system.

- Michelle Donelan

U.K. secretary of state for science

1.1 Motivation

The shortage of medical resources poses a significant global challenge. According to reports, approximately 47% of the global population lacks access to adequate diagnostic services (Fleming et al., 2021). This deficiency not only complicates the provision of timely and essential medical care but also adversely impacts the quality of life and well-being of the general population.

In the 21st century, the healthcare industry is undergoing a digital transformation, marked by a shift toward intelligent healthcare (Mathews et al., 2019; Turner et al., 2019). Fuelled by advancements in artificial intelligence (AI), including machine learning (ML) and, especially, deep learning (DL), the automation of medical diagnostics and healthcare delivery demonstrates tremendous potential. This offers a concrete solution to alleviate the strain on clinical resources and manpower (Reddy et al., 2020; Rajpurkar et al., 2022; Wang and Preininger, 2019).

Physiological data consist of measurements or recordings of various bodily functions and processes (Inbamani et al., 2022). Such data include information about vital signs, activities, and bodily responses, primarily collected in clinical settings. They play a crucial role in health monitoring and diagnostics. One typical example of physiological data is electrocardiogram (ECG) signals. These signals record the electrical activity of the heart, including its rate and rhythm, which are used for monitoring heart rhythm and identifying cardiovascular abnormalities (Yang et al., 2020). Training doctors capable of analysing physiological data takes years, while the

use of AI and ML can significantly enhance the accuracy and efficiency of analysing these data, reducing the reliance on manual processes. For example, research has demonstrated that AI-enhanced ECGs, acquired during normal sinus rhythm, can identify individuals with atrial fibrillation at the point of care (Hygrell et al., 2023). Furthermore, advancements in mobile sensing technologies mark a significant development, enabling the collection of various physiological data via mobile devices and wearables (Steinhubl et al., 2015). This progress paves the way for delivering healthcare services anytime and anywhere.

In light of this, this thesis is dedicated to exploring machine learning, with a particular focus on deep learning, for the analysis and modelling of physiological data. Specifically, it delves into the modelling of respiratory audio for respiratory health screening, ECG signals for predicting cardiovascular diseases, and dermoscopic images for detecting skin cancer, as detailed in Chapter 2. This research utilises data collected both in clinical settings and through mobile devices. The ultimate aim is to develop cost-effective and efficient health diagnostics that are accessible to a broad population.

Nevertheless, the development of high-performing ML models for physiological data is a non-trivial task. Similar to many other applications, the performance of these models heavily depends on the quality and quantity of available physiological data for model parameter learning, as well as the effectiveness of learning strategies, ultimately determining the reliability of the models (Jordan and Mitchell, 2015; LeCun et al., 2015; Litjens et al., 2017). The data challenges impeding our goal are limited access to comprehensive and diverse physiological datasets, data biases arising from underrepresented demographics and specific healthcare settings, and constraints in data collection methods. Moreover, ethical considerations and privacy concerns further complicate the data collection process. Consequently, available physiological datasets associated with various specific diseases are often small in size, noisy, skewed towards certain health conditions, and distributed across multiple sources. These factors may lead to under-performing machine learning models and overconfident diagnoses. A detailed discussion of how these factors present substantial obstacles to our goal is elaborated in Chapter 1.2.

To overcome these challenges, this thesis endeavours to develop novel and cutting-edge ML and DL solutions for physiological data. By providing insights into healthcare applications, these studies aim to pave the way for practical, reliable, and cost-effective AI-empowered health diagnostics.

1.2 Challenges and research questions

Developing high-performing machine learning models for health diagnostics in real-world scenarios poses numerous challenges. This thesis specifically targets challenges arising from the

limited availability of data and the paradigms of model training. Our objective is to offer meaningful solutions that enhance the reliability and effectiveness of automated health diagnostics through the use of physiological data.

i) Insufficient and imbalanced physiological data for model development. The effectiveness of a deep learning model largely depends on its complexity, which in turn is determined by the number of parameters it contains. Supervised learning becomes a critical approach in this context, because it directly leverages the relationship between input data and corresponding labels to effectively train these complex models, although it does not preclude the use of alternative methods, such as self-supervised or semi-supervised learning (Cho et al., 2015). Gathering extensive physiological data across diverse health conditions for machine learning research presents challenges. These challenges can stem either from the low prevalence of certain diseases, limiting the pool of volunteers who can contribute data, or from the need for clinical verification of these health conditions, which is often time-consuming and financially burdensome. Moreover, available labelled physiological data for research frequently exhibit significant class imbalance (Goldberger et al., 2000). For instance, the ICBHI 2017 respiratory audio database consists of a total of 5.5 hours of recordings containing 6,898 respiratory cycles, of which 1,864 contain crackles (27.0%), 886 contain wheezes (12.9%), and 506 contain both crackles and wheezes (7.3%) (Rocha et al., 2019). Training on such insufficient and class-imbalanced data can cause the deep learning model to disproportionately prioritise the majority class, resulting in subpar performance on the minority class and, consequently, sub-optimal disease detection outcomes.

ii) Deep learning models can produce overconfident predictions. Despite the remarkable performance on testing sets with distributions identical to the training data, these models often struggle to capture the inherent uncertainty arising from environmental factors, data variability, and training methodologies. As a result, these models may exhibit overconfidence when deployed in the real physical world (Gal and Ghahramani, 2015; Louizos and Welling, 2017; Ovadia et al., 2019). In certain situations, such as when input data are of low quality, deviate from the training set, or involve input categories not encountered during training, these models may generate predictions with unwarranted confidence. For instance, a model trained to recognise cats and dogs might confidently misclassify rabbits as cats. Similarly, a model designed to distinguish asthma patients from healthy individuals might erroneously categorise individuals with lung cancer as asthma patients, as lung cancer falls outside the model's training categories. This issue becomes particularly severe when deep learning is optimised with limited and imbalanced physiological data for health diagnostics, leading to overly confident yet incorrect diagnoses (Park et al., 2021). Such overconfidence in healthcare poses a significant risk and could result in severe consequences. For example, when a model confidently classifies a cancer patient into a non-cancer category, the high confidence level may lead clinicians to assign a lower priority to checking and correcting the model's prediction. Consequently, the patient may miss the opportunity for

timely treatment. Hence, it is imperative to devise suitable training methods and tackle the issue of overconfident diagnoses to effectively mitigate these risks.

iii) Health data can be privacy-sensitive, and thus problematic for sharing with model developers. Typically, machine learning research in terms of developing models for health diagnostics requires gathering physiological data and health conditions from individuals. These data are then transmitted to a central server where the model parameters can be optimised. However, this approach raises privacy concerns since personal health information is highly sensitive and should be protected from unauthorised access. A traditional solution is anonymously sharing data and the usage of the shared data is subject to certain restrictions. However, such a process can slow down data collection and model development (Crow et al., 2006; Kreuter et al., 2020), and does not eliminate the risk of malicious attacks occurring during data transmission to or storage on the server (Li and Liu, 2021). To enhance privacy protection, it is essential to design model training methods that allow the data to remain at their original location, instead of transferring them to a central server. Although many decentralised learning approaches have been proposed, their performance is hindered by the class imbalance and heterogeneous distributions of physiological data residing across personal mobile devices or health institutions. Effective solutions are needed to address these problems.

Given the above discussion, the central research questions (RQs) guiding the thesis are summarised below:

- ***RQ 1****: How can we mitigate the bias and calibrate the confidence of predictions when training models for health screening with limited and imbalanced physiological data?*
- ***RQ 2****: How can we develop high-performing models with efficient uncertainty estimation for health diagnostics, given multi-class imbalanced physiological data?*
- ***RQ 3****: How can we train deep learning-driven health screening models using only decentralised and imbalanced physiological data stored on mobile devices?*
- ***RQ 4****: How can we develop high-performing models for health diagnostics using physiological data distributed in multiple places and with heterogeneous distributions?*

Aligned with addressing these research questions, this thesis introduces several innovative techniques for physiological data. These solutions are designed to improve the reliability of automated health diagnostics and mitigate the privacy risk of model development, thereby facilitating the practical deployment of deep learning in real-world scenarios.

In summary, the primary objective of this thesis is to harness ***limited and imbalanced physiological data*** *to develop deep learning models capable of providing* ***reliable automated health diagnostics*** *while* ***minimising privacy risks in model training****.*

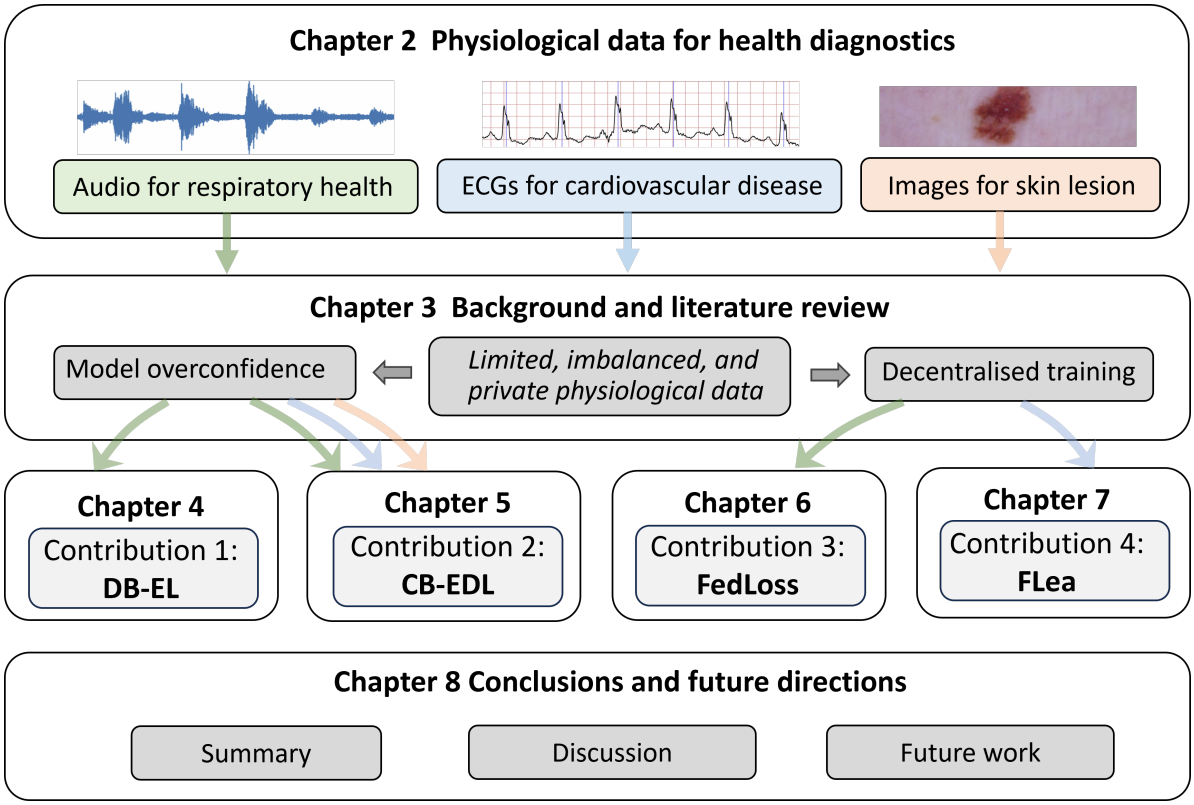


Figure 1.1: **Outline of the remaining chapters of this thesis**.

1.3 Contributions and chapter outline

The structure of this thesis is depicted in Figure 1.1. In terms of methods, we tailor deep learning models to suit the unique properties of physiological data. In terms of applications, we explore complex challenges in health diagnostics. We begin with an introduction to the physiological data and clinical tasks addressed in this thesis, presented in Chapter 2. Following a review of related work in Chapter 3, we detail our four primary contributions. Chapter 8 provides the conclusion of the thesis. The main contributions include:

Contribution 1: A calibrated and uncertainty-aware deep learning method *DB-EL* using limited and imbalanced physiological data for binary health screening.

To address the challenges posed by model overconfidence and bias resulting from limited and imbalanced physiological data, in Chapter 4, we propose an innovative data-balanced ensemble learning framework (*DB-EL*) for binary health screening. This framework tackles the issue of class imbalance by re-sampling from the healthy class, creating balanced data subsets for training multiple deep learning models. The outputs of these models are then fused to provide the final calibrated health diagnoses. The inconsistency of the predictions made by those models is utilised as a measurement of model uncertainty to indicate the correctness of each diagnosis.

We evaluate this approach using a case study involving a physiological audio-driven COVID-19 screening application. In this scenario, the development of the model is challenged by a limited number of available audio samples, along with a significant imbalance in the count of COVID-19 positive and negative samples.

Our experimental findings reveal that our method outperforms baseline methods in terms of screening accuracy and confidence calibration. Moreover, we find that highly uncertain model predictions tend to correspond to incorrect diagnoses. This suggests that our quantified uncertainty can enhance the reliability of automated health screening by recognising the uncertain diagnoses.

Contribution 2: An accurate and efficient uncertainty quantification method *CB-EDL* for deep learning with multi-class imbalanced physiological data.

Recognising the significance of model calibration in Chapter 4, we further investigate the practical use of uncertainty estimation for mobile health applications. The ensemble learning approach proposed in Chapter 4 is confined to binary health screening and proves inefficient, requiring multiple models and multiple forward passes. In contrast, evidential deep learning (EDL) leverages distribution-based rather than deterministic outputs, enabling both model and data uncertainty quantification through a single model, thereby enhancing efficiency. However, it is susceptible to data bias in the presence of class imbalance.

To render this efficient EDL method effective on multi-class imbalanced physiological data, Chapter 5 introduces a class-balanced EDL (*CB-EDL*) approach with two novel mechanisms tailored for handling class imbalance. EDL transforms the learned classification evidence into a Dirichlet distribution, from which the uncertainty can be efficiently estimated. Our proposed class-balanced EDL enhances the vanilla EDL by *i)* a class-level pooling loss to mitigate the bias in classification evidence; and *ii)* a learnable prior that is regularised by the class distribution to facilitate learning for minority classes. This approach can overcome the unsuitable assumption that data is uniformly distributed across classes in the original EDL theory.

The superiority of our method is validated through clinical tasks using three real-world physiological datasets and a general classification task using machine learning benchmark data. In comparison with long-tailed learning baselines designed for class imbalance, our method exhibits more calibrated predictions. When compared to other uncertainty quantification counterparts, our method proves to be more efficient while maintaining superior performance. Moreover, we highlight the utility of quantified predictive uncertainty in healthcare applications: it facilitates the detection of misclassification and out-of-training distributed instances, thereby substantially reducing the risk of misdiagnoses.

Contribution 3: A federated learning approach *FedLoss* to develop health screening models using imbalanced physiological data distributed across mobile devices.

Chapter 6 explores the feasibility of training deep learning models for health screening without the need to aggregate physiological data from mobile devices. The chapter delves into the realm of Federated Learning (FL), where mobile devices independently train models using their private data, and only the model parameters are synchronised and aggregated into a global model on the server.

A practical challenge in this scenario is data imbalance: each participant, *i.e.*, a mobile device holder, represents only a single health status, and globally, there are usually more healthy participants than unhealthy ones. Such data distribution could lead to biased federated model aggregation. To improve the model effectiveness while ensuring data privacy, this chapter proposes a weighted model parameter aggregation method, named *FedLoss*.

The proposed method is validated based on the physiological audio-driven COVID-19 detection task by simulating the real-world cross-device setting. Experimental results indicate that the model's performance is comparable to that of a model trained on centralised data. This work opens the door to privacy-preserving mobile health research by turning data aggregation into model aggregation.

Contribution 4: A federated learning approach *FLea* to develop deep learning models for health diagnostics using distributed and heterogeneous physiological data.

Building upon the study presented in Chapter 6, Chapter 7 explores another *cross-silo* federated learning setting. In this scenario, multiple health institutes, such as hospitals, possess physiological data from various health conditions. Functioning as independent data silos, these institutes collaborate on the development of a deep learning model for health diagnostics without exchanging raw data. However, the diverse prevalence rates of diseases in these institutes result in data heterogeneity, posing a distinctive challenge to the performance of federated learning.

To address this challenge, Chapter 7 introduces *FLea*, a novel feature sharing and augmentation method to model heterogeneous physiological data distributed in multiple sites. Health institutes utilising *FLea* not only share model parameters but also exchange privacy-protected features, specifically intermediate layer activations from the model, to enhance local training. The utilisation of these features alleviates the local model drift caused by data heterogeneity, ultimately enhancing the global model's performance upon aggregation.

To evaluate *FLea*, our experiments leverage multi-centre ECG data where each data silo contains only a subset of identified cardiac arrhythmia cases. Notably, *FLea* achieves competitive accuracy compared to a model trained with centralised data. Furthermore, we evaluate *FLea*

by distributing machine learning benchmark data into multiple silos with varying levels of label skewness. The results demonstrate its generally superior performance over other FL counterparts.

1.4 List of publications

The research outcomes presented in this thesis have resulted in several publications and submissions at renowned machine learning and signal processing conferences, as well as health-focused journals. Throughout the peer review and presentation process, the valuable feedback received greatly contributed to the development of solid research ideas and the formulation of this thesis.

In collaboration with other researchers in the Mobile Systems group, a physiological audio database for respiratory health research was published at NeurIPS dataset and benchmark track [1]. Additionally, a deep learning model tailored for this physiological audio database was proposed, as detailed in NPJ Digital Medicine [2]. Both this dataset and model serve as consistent evaluation tools for the proposed methods in Chapters 4 and 6. Furthermore, the methodology presented in Chapter 4 is derived from a paper presented at INTERSPEECH [3]. Chapter 5 is based on a paper presented at the Workshop at NeurIPS [4], which has been extended and published in JBHI [5]. Chapter 6 builds upon a recently published paper at ICASSP [6], while Chapter 7 is a work that has been submitted to KDD [7].

In addition, I have collaborated on publications in the broader field of deep learning and data science for public health. While these works are not directly related to this thesis, they have significantly influenced my ideas and contributed to the enhancement of my research skills.

Works related to and covered by this thesis (* *equal contribution*):

1. **Xia, T.***, Spathis, D.*, Ch, J., Grammenos, A., Han, J., Hasthanasombat, A., ... & Mascolo, C. (2021). COVID-19 sounds: A large-scale audio dataset for digital respiratory screening. *In Proceedings of the Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track, NeurIPS 2021* (Xia et al., 2021d) (*2nd Poster Award in Precision Health Initiative Launch Symposium, Cambridge, 2022*).

2. Han, J.*, **Xia, T.***, Spathis, D., Bondareva, E., Brown, C., Chauhan, J., ... & Mascolo, C. (2022). Sounds of COVID-19: Exploring realistic performance of audio-based digital testing. *NPJ Digital Medicine* (Han et al., 2022).

3. **Xia, T.**, Han, J., Qendro, L., Dang, T., & Mascolo, C. (2021). Uncertainty-aware COVID-19 detection from imbalanced sound data. *In Proceedings of the 22nd Annual Conference of the International Speech Communication Association, INTERSPEECH 2021* (Xia et al., 2021a).

4. **Xia, T.,** Han, J., Qendro, L., Dang, T., & Mascolo, C. Hybrid-EDL: Improving evidential deep learning for uncertainty quantification on imbalanced data. *In Workshop on Trustworthy and Socially Responsible Deep Learning, NeurIPS 2022* (Xia et al., 2022c).

5. **Xia, T.**, Dang, T., Han, J., Qendro, L., & Mascolo, C. (2023) Uncertainty-aware health diagnostics via class-balanced evidential deep learning. *IEEE Journal of Biomedical and Health Informatics, JBHI 2024* (Xia et al., 2024a).

6. **Xia, T.**, Han, J., Ghosh, A., & Mascolo, C. (2023). Cross-device federated learning for mobile health diagnostics: A first study on COVID-19 detection. *In Proceedings of the ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2023* (Xia et al., 2023a).

7. **Xia, T.**, Ghosh, A., & Mascolo, C. FLea: Addressing data scarcity and label skew in federated learning via privacy-preserving feature augmentation. (2024). *In Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery & Data Mining (KDD), KDD 2024* (Under review) (Xia et al., 2024b).

Other publications beyond this thesis (*in chronological order*):

- Brown, C.*, Chauhan, J.*, Grammenos, A.*, Han, J.*, Hasthanasombat, A.*, Spathis, D.*, **Xia, T.***, Cicuta, P., & Mascolo, C. (2020). Exploring automatic diagnosis of COVID-19 from crowdsourced respiratory sound data. *In Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2020* (Brown et al., 2020a) (*Better Further Award by the Department of Computer Science and Technology, University of Cambridge, 2021*).

- Han, J., Brown, C.*, Chauhan, J.*, Grammenos, A.*, Hasthanasombat, A.*, Spathis, D.*, **Xia, T.***, Cicuta, P., & Mascolo, C. (2021). Exploring automatic COVID-19 diagnosis via voice and symptoms from crowdsourced data. *In Proceedings of the ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2021* (Han et al., 2021a).

- **Xia, T.**, Qi, Y., Feng, J., Xu, F., Sun, F., Guo, D., & Li, Y. (2021). Attnmove: History enhanced trajectory recovery via attentional network. *In Proceedings of the AAAI Conference on Artificial Intelligence, AAAI 2021* (Xia et al., 2021c).

- **Xia, T.**, Lin, J., Li, Y., Feng, J., Hui, P., Sun, F., ... & Jin, D. (2021). 3DGCN: 3-dimensional dynamic graph convolutional network for citywide crowd flow prediction. *ACM Transactions on Knowledge Discovery from Data, TKDD 2021* (Xia et al., 2021b).

- Ghosh, A.*, & **Xia, T.***. (2021). Mobility-based individual POI recommendation to control the COVID-19 spread. *In Proceedings of the 2021 IEEE International Conference*

on Big Data, Big Data 2021 (Ghosh and Xia, 2021).

- **Xia, T.**, Han, J., & Mascolo, C. (2021). Benchmarking uncertainty quantification on biosignal classification tasks under dataset shift. *In Multimodal AI in healthcare: A paradigm shift in health intelligence* (Xia et al., 2022a).

- **Xia, T.**, Han, J., & Mascolo, C. (2022). Exploring deep learning for audio-based respiratory condition screening: A concise review of databases, methods, and open issues. *Experimental Biology and Medicine, EBM 2022* (Xia et al., 2022b).

- Li, T., **Xia, T.**, Wang, H., Tu, Z., Tarkoma, S., Han, Z. & Hui, P., (2022). Smartphone app usage analysis: Datasets, methods, and applications. *IEEE Communications Surveys & Tutorials, 2022* (Li et al., 2022).

- Feng, T., **Xia, T.**, Fan, X., Wang, H., Zong, Z., & Li, Y. (2022). Precise mobility intervention for epidemic control using unobservable information via deep reinforcement learning. *In Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD 2022* (Feng et al., 2022b).

- Bondareva, E., **Xia, T.**, Han, J., & Mascolo, C. (2022). Towards uncertainty-aware murmur detection in heart sounds via tandem learning. *In Proceedings of the 2022 Computing in Cardiology, CinC 2022* (Bondareva et al., 2022).

- Dang, T., Han, J.*, **Xia, T.***, Spathis, D., Bondareva, E., Siegele-Brown, C., ... & Mascolo, C. (2022). Exploring longitudinal cough, breath, and voice data for COVID-19 progression prediction via sequential deep learning: model development and validation. *Journal of Medical Internet Research, JMIR 2022* (Dang et al., 2022).

- **Xia, T.**, Li, Y., Qi, Y., Feng, J., Xu, F., Sun, F., Guo, D., & Jin, D. (2023) History-enhanced and uncertainty-aware trajectory recovery via attentive neural network. *ACM Transactions on Knowledge Discovery from Data, TKDD 2023* (Xia et al., 2023b).

- Dang, T., Han, J.*, **Xia, T.***, Bondareva, E., Siegele-Brown, C., Chauhan, J., Cicuta, P., & Mascolo, C. (2023). Conditional neural ODE processes for individual disease progression forecasting: a case study on COVID-19. *In Proceedings of the 28th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2023* (Dang et al., 2023).

- Zhang, Y., **Xia, T.**, Ghosh, A., & Mascolo, C. (2023). Uncertainty quantification in federated learning for heterogeneous health data. *In International Workshop on Federated Learning for Distributed Data Mining, co-located with the 29th ACM SIGKDD Conference, KDD 2023* (Zhang et al., 2023d).

- Han, Z., **Xia, T.**, Xi, Y., & Li, Y. (2023). Healthy Cities, A comprehensive dataset for environmental determinants of health in England cities. *Scientific Data* (Han et al., 2023b).
- Feng, T., Song, S., **Xia, T.**, & Li, Y., (2023). Contact tracing and epidemic intervention via deep reinforcement learning. *ACM Transactions on Knowledge Discovery from Data, TKDD 2023* (Feng et al., 2023).
- Han, J., Montagna, M., Grammenos, A., **Xia, T.**, Bondareva, E., Siegele-Brown, C., ... & Mascolo, C. (2023). Evaluating Listening Performance for COVID-19 Detection by Clinicians and Machine Learning: Comparative Study. *Journal of Medical Internet Research (JMIR)* (Han et al., 2023a).
- Yfantidou, S., Spathis, D., Constantinides, M., **Xia, T.**, & Van Berkel, N., (2023). FairComp: Workshop on fairness and robustness in machine learning for ubiquitous computing. *In Adjunct Proceedings of the 2023 ACM International Joint Conference on Pervasive and Ubiquitous Computing & the 2023 ACM International Symposium on Wearable Computing* (Yfantidou et al., 2023).
- Zong, Z., **Xia, T.**, Zheng, M. and Li, Y., (2024). Reinforcement Learning for Solving Multiple Vehicle Routing Problem with Time Window. *ACM Transactions on Intelligent Systems and Technology, TIST 2023* (Zong et al., 2024).

Chapter 2

Physiological data for health diagnostics

Biology can be divided into the study of proximate causes, the study of the physiological sciences, and into the study of ultimate causes, the subject of natural history.

- Ernst Mayr

German-American evolutionary biologist

2.1 Introduction to physiological data

2.1.1 Overview

Physiological data refers to measurements or recordings of various bodily functions and processes (Inbamani et al., 2022). These data encompass information about the body's vital signs, activities, and responses. Common examples of physiological data include: *electrocardiogram signals (ECGs)*, which measure the electrical activity of the heart; *electroencephalogram signals (EEGs)*, which record electrical activity in the brain; *electromyogram signals (EMGs)*, which monitor electrical activity produced by skeletal muscles; *blood pressure signals*, which reflect the force of blood as it moves through the arteries; and *respiration signals*, which indicate the rate and depth of breathing Bhatt et al. (2021). The wealth of such data can provide healthcare professionals with valuable information for making informed diagnostics and preventive decisions (Orphanidou, 2019; Rim et al., 2020).

Traditionally, the collection of physiological data has required clinical devices operated by medical professionals. However, with advancements in mobile sensing and wearable technology, it is now possible to gather some physiological data using mobile devices, facilitating ambulatory monitoring. Examples of such data include: *photoplethysmogram signals (PPGs)*, which

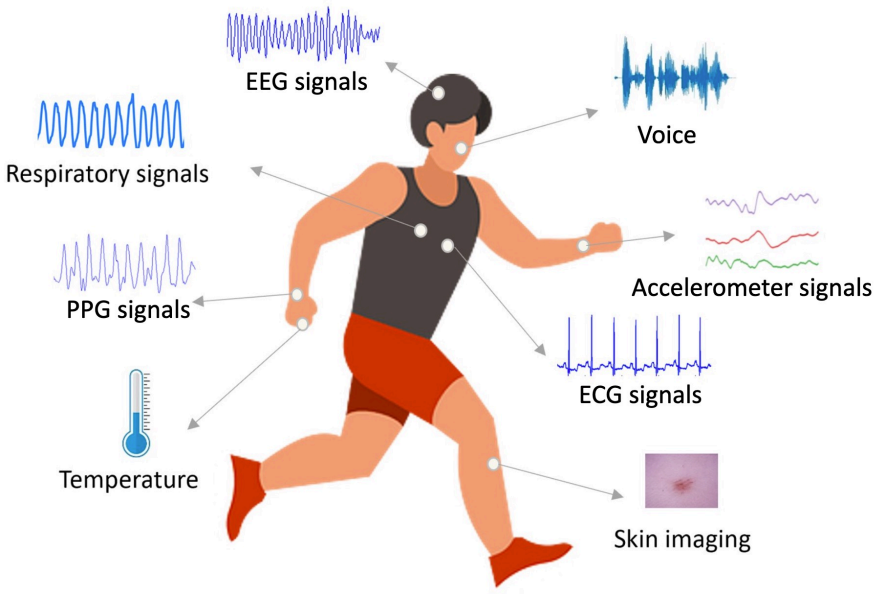


Figure 2.1: **An example of physiological data**. Sensors can be attached to different parts of the human body to collect various physiological data.

measure blood volume changes in the microvascular bed of tissue; *respiratory audio*, comprising recordings of sounds produced by the respiratory system during breathing, captured using microphones; *dermoscopic images*, a type of photograph taken using mobile dermoscopy in dermatology; and movements, gestures, and overall motion patterns, recorded by *inertial measurement units (IMUs)*, which offer insights into physical activity and step counts. The emergence of mobile health devices has significantly expanded the availability of physiological data collection.

An example of different physiological data is illustrated in Figure 2.1. Physiological data are characterised by their dynamic nature, capturing changes over time and providing real-time insights into health parameters. However, this type of data is usually susceptible to noise and interference from external factors, necessitating pre-processing to enhance accuracy. Privacy considerations are crucial due to the sensitive nature of health information, requiring secure storage and transmission. Additionally, integration with other health metrics, such as symptoms and medical histories, is common, providing a holistic understanding of an individual's well-being. Managing these characteristics ensures effective interpretation and utilisation of the wealth of physiological data for health monitoring and diagnostics.

The aim of this thesis is to use machine learning for modelling physiological data to diagnose health conditions. Machine learning is a paradigm that employs functions and sets of parameters to model data and predict future outcomes Jordan and Mitchell (2015). To adjust the model parameters accurately, physiological data along with corresponding health condition labels are required. Further details on model training are introduced in Chapter 3.1. Chapter 2.1.2 dis-

cusses the process of data collection and annotation.

2.1.2 Data collection and pre-processing

Physiological data can be collected and annotated either simultaneously or consecutively. In the former approach, physiological signals are gathered from a specific group of individuals whose health conditions have been previously examined. For instance, when investigating the decline in mobility and language ability caused by dementia, researchers study both dementia patients and healthy controls. They gather IMU data and speech samples from each group, respectively (Syed et al., 2020; Carissimo et al., 2023). In some studies, physiological data are collected from the population to a central location and then annotated by clinicians. An example of this approach is the SAFER study, where ECG signals are collected from a large population and then screened and labelled by cardiologists to identify conditions such as Atrial Fibrillation (Akande et al., 2023; Hygrell et al., 2023). These annotations play a crucial role in supervising the model learning process, as further discussed in Chapter 3.1.2.

Pre-processing physiological data samples before inputting them into a machine learning model is also necessary, which typically involves several steps (Nabian et al., 2017; Sajno et al., 2023; Xia et al., 2022b). The initial step usually includes data cleaning, which encompasses removing any noise, artifacts, or outliers from the data. This process may involve filtering out noise, correcting signal abnormalities, or eliminating data points considered erroneous. Additionally, normalising the data can be beneficial to the modelling by scaling them to a standardised range. This ensures that all features are on a similar scale, facilitating faster convergence of the model during training (Koh, 2019). Furthermore, additional pre-processing steps, such as dimensional reduction, data augmentation (generating additional training samples through transformations or perturbations), and temporal alignment, are specific to the application at hand. Determining the most suitable pre-processing steps often requires experimentation and domain expertise.

2.2 Medical tasks and physiological datasets

In this thesis, we consider a variety of physiological datasets encompassing different data modalities for conducting experiments. The following sections offer a concise overview of the specific medical tasks and datasets utilised.

2.2.1 Physiological audio data for respiratory health screening

Digital audio is an informative and easy-to-collect modality for health status monitoring (Mascolo, 2020). Recently, researchers have started to explore whether respiratory sounds could be used for the diagnosis of COVID-19 (Deshpande and Schuller, 2020). Stethoscope data from

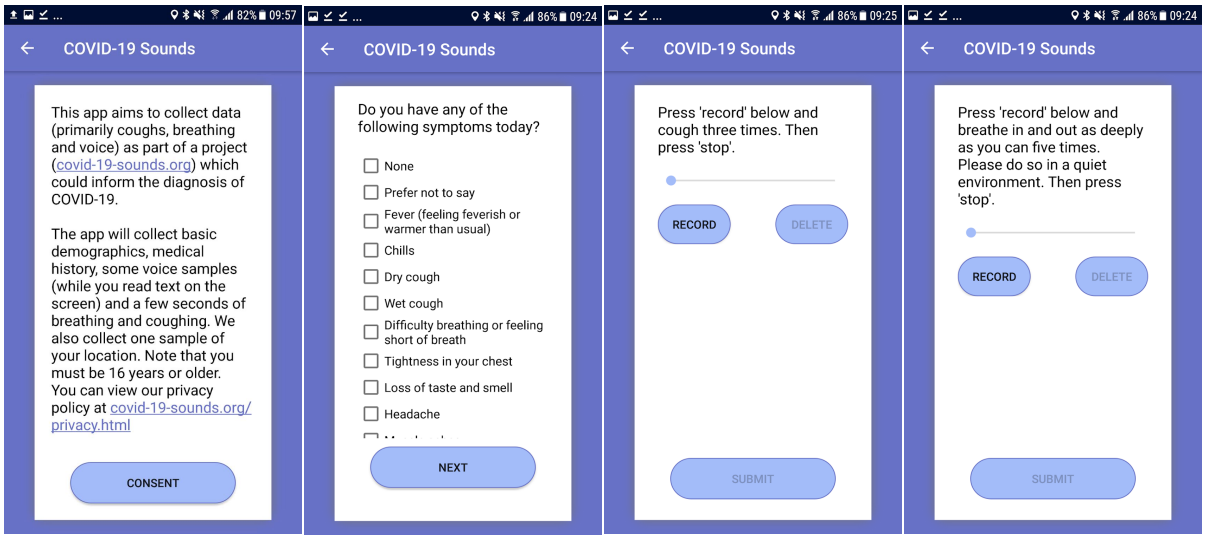


Figure 2.2: **Screens of the COVID-19 Sounds data collection app.** The users are asked to input their symptoms along with medical history, as well as to record breathing, cough, and voice sounds every couple of days.

lung auscultation (Huang et al., 2020), coughs collected by phones (Imran et al., 2020), and speech recordings (Han et al., 2020) have been analysed to distinguish COVID-19 patients from healthy participants. To explore the power of sounds in mobile health applications, two audio datasets are leveraged for experiments in this thesis.

COVID-19 Sounds data. This data was collected through the COVID-19 Sounds App[1], and it received approval from the Ethics Committee of the Department of Computer Science and Technology at the University of Cambridge. Participants in the data collection were asked to provide information about their demographics, medical history, and smoking status. Additionally, they were required to disclose their COVID-19 test results, hospitalisation status, and any symptoms they experienced. To capture respiratory sounds, participants were instructed to: *i)* Cough three times; *ii)* Breathe deeply through their mouth three to five times; and *iii)* Read a short sentence on-screen and repeat it three times. Figure 2.2 displays some screenshots of the App (Android version).

After a year of collecting data via the app, a total of 36,116 participants worldwide contributed to the database, resulting in 53,449 audio samples (totalling over 552 hours) (Xia et al., 2021d). Despite the substantial number of samples, the majority remain unlabelled, meaning they lack confirmed COVID-19 status. Specifically, only 1,139 samples reported a positive COVID-19 test result (within or before the last 14 days), while 5,251 samples reported a negative result. This database is employed for various purposes in this thesis, with subsets utilised for the studies

[1] https://www.covid-19-sounds.org/en/

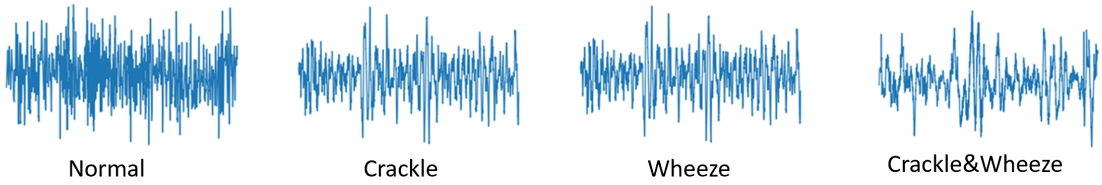


Figure 2.3: **Examples of ICBHI challenge data.** These audio samples are of different lengths (from 20ms to 100ms) and are associated with different respiratory abnormalities.

outlined in Chapters 4.3 and 6.3, respectively.

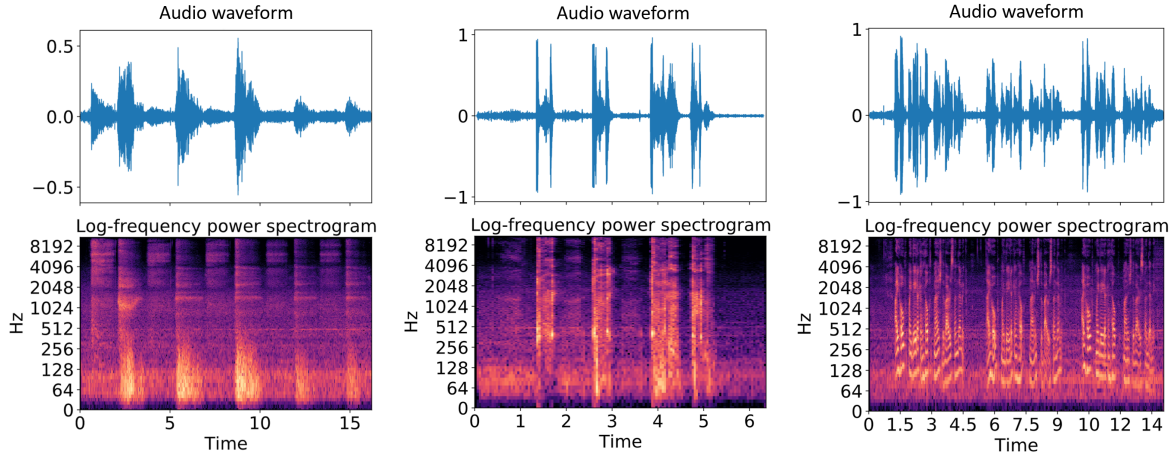
It is noteworthy to highlight my contribution to the management of this dataset during my PhD. This *COVID-19 Sounds* data represents the most extensive multi-modal collection of respiratory sounds, covering three modalities: breathing, cough, and voice recordings. We presented the characteristics of the data and established a benchmark for modelling it at NeurIPS 2021 (Xia et al., 2021d). Importantly, we have made the data accessible to more than 400 research institutes. Throughout my PhD, I consistently utilised this database as a crucial resource to validate my work.

ICBHI challenge data. ICBHI 2017 Respiratory Challenge[2] published a dataset collected from multiple microphones and stethoscopes (Rocha et al., 2019). This respiratory sound database contains audio samples, collected independently by two research teams in two different countries, over several years. Most of the database consists of audio samples recorded by the School of Health Sciences, University of Aveiro (ESSUA) research team at the Respiratory Research and Rehabilitation Laboratory (Lab3R), ESSUA, and at Hospital Infante D. Pedro, Aveiro, Portugal. The second research team, from the Aristotle University of Thessaloniki (AUTH) and the University of Coimbra (UC), acquired respiratory sounds at the Papanikolaou General Hospital, Thessaloniki and at the General Hospital of Imathia (Health Unit of Naousa), Greece.

The database consists of a total of 5.5 hours of recordings containing 6,898 respiratory cycles, of which 1,864 contain crackles (27.0%), 886 contain wheezes (12.9%), and 506 contain both crackles and wheezes (7.3%), in 920 annotated audio samples from 126 subjects. The frequency range of healthy vesicular breathing lung sounds extends up to 1,000 Hz, where the majority of the spectrum power falls within the range from 60 to 600 Hz. The dataset is utilised to evaluate the performance of multi-class respiratory abnormality detection, as presented in Chapter 5.3. An example is given in Figure 2.3.

Spectrograms. Audio data are usually collected with a high sampling rate (above kHz), making the direct analysis of waveform challenging. Signal processing techniques are important in this

[2] https://bhichallenge.med.auth.gr/



(a) Breathing recording. (b) Cough recording. (c) Voice recording.

Figure 2.4: **Examples of COVID-19 Sounds data**. An example of recordings' waveforms and the associated spectrograms from a participant who tested COVID-19 positive within 14 days of the recording. The participant is a male aged over 30, with a smoking history, speaks English, and had symptoms including a wet cough, headache, and sore throat on the recording day.

context to transform the data into a more representative format for subsequent analysis. One typical technique is the Short-Time Fourier Transformation (STFT), which generates spectrograms, *i.e.*, two-dimensional representations of a one-dimensional time series sample.

As a variation of the Fourier Transform, the STFT provides a time-dependent representation of the frequency components of an audio signal (Oppenheim, 1999). Given a mono-channel audio sample x, the STFT is defined by the energy coefficient at any time t and frequency f. Since audio samples are discrete digital signals, it is common to use the discrete STFT. Specifically, x obtained by sampling frequency f_s will be segmented into overlapped short segments through window sliding, *i.e.*, each segment has a duration T containing N data points with $N = T \cdot f_s$. The frequencies are considered by frequency bins, which are evenly spaced between 0 Hz and the Nyquist frequency (*i.e.*, half of f_s). Mathematically, the discrete STFT matrix is,

$$X[n,k] = \sum_{m=0}^{N-1} x[n+m] \cdot w[m] \cdot e^{-j2\pi km/N}, \tag{2.1}$$

where $w[m]$ is the value of the window function at point m, and $e^{-j2\pi km/N}$ represents the complex sinusoidal basis function at frequency bin k for point m in this segment.

The commonly used window function in signal processing, and specifically in the context of STFT for audio, is the Hann window (Harris, 1978). It is a type of tapering window that gradually decreases towards the edges. It is defined mathematically as,

$$w[m] = 0.5 - 0.5\cos\left(\frac{2\pi m}{N-1}\right). \tag{2.2}$$

The Hann window is popular because it offers a good compromise between the main lobe width and the side lobe attenuation, resulting in reduced spectral leakage and improved frequency resolution compared to other window functions. It is often used in applications such as spectral analysis, audio processing, and speech recognition.

The widely used spectrograms are the *power spectrograms*, which represent the magnitude of the complex STFT coefficient (Hatamian et al., 2020). This choice is made due to its ability to convey information regarding the power (or energy) distribution across various frequency components and time intervals within a given signal. A power spectrogram for x can be derived by:

$$P[n, k] = |X[n, k]|^2. \tag{2.3}$$

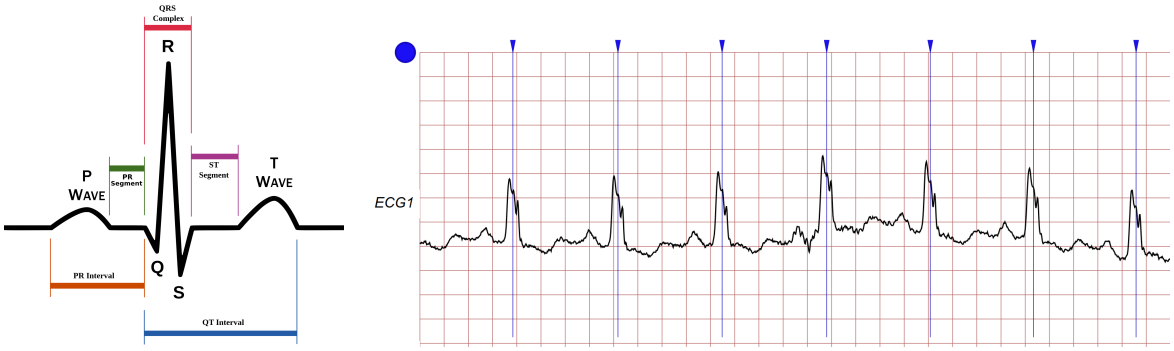
Examples of power spectrograms of respiratory audio samples are shown in Figure 2.4. For this example, we slide the window by every 1024 data points to segment the audio, with a segment length $N = 2048$ given the sampling rate of 22.05 kHz. Hann window is used. The obtained spectrograms reflect good temporal-frequency dynamics for subsequent feature extraction. In this thesis, we will consistently use spectrograms as the representation for audio.

2.2.2 Electrocardiogram data for cardiovascular disease prediction

Cardiovascular disease is the leading cause of death worldwide (Tsao et al., 2023). Early treatment can prevent serious cardiac events, and the most important tool for screening and diagnosing cardiac electrical abnormalities is through electrocardiograms (ECG) (Kligfield et al., 2007). The ECG is a noninvasive representation of the electrical activity of the heart that is measured using electrodes placed on the torso. An ECG signal of a heartbeat is illustrated in Figure 2.5(a)), showing the length of time it takes for the initial impulse to fire and then ends in the contracting of depolarisation, a process in which the electrical charge of heart cells is reset to allow for the next heartbeat. Figure 2.5(b) presents a single-channel ECG recording, allowing observation of the regularity or irregularity of the heart's rhythm. Normal sinus rhythm is regular, while irregularities can indicate various conditions, such as arrhythmia.

ECG5000 data. We first introduce a pro-processed single-channel ECG database *ECG5000* for experiments. This is an ECG database that includes long-term ECG recordings from 15 subjects (11 men, aged 22 to 71, and 4 women, aged 54 to 63) with severe congestive heart failure (NYHA class 3–4)[3]. This group of subjects was part of a larger study group receiving conventional medical therapy before receiving the oral inotropic agent, milrinone. The individual recordings are each about 20 hours in duration and contain two ECG signals each sampled at 250 samples per second with 12-bit resolution over a range of ±10 millivolts. The original analogue

[3] https://www.physionet.org/content/chfdb/1.0.0/



(a) Hearbeat. (b) ECG.

Figure 2.5: **ECG data samples**. (a) presents a schematic diagram of normal sinus rhythm for a human heart as seen on ECG. (b) shows a single-channel ECG recording from a normal individual.

recordings were made at Boston's Beth Israel Hospital (now the Beth Israel Deaconess Medical Center) using ambulatory ECG recorders with a typical recording bandwidth of approximately 0.1 Hz to 40 Hz. To ease model learning, this database has been further split and interpolated into 5000 equal-length (140) heartbeats[4]. It consists of five classes: 58.4% are normal, 35.3% have heat failure typed R-on-T premature ventricular contraction, 3.9% premature ventricular contraction, 2.0% supraventricular premature or ectopic beat, and 0.5% unclassified beats (Chen et al., 2015). This dataset is leveraged for heart failure prediction. Also because of its severe imbalance character, this data is included in the study in Chapter 5.3.

CIC2020 data. The standard 12-lead ECG is widely used for diagnosing various cardiac arrhythmias, such as atrial fibrillation, and other cardiac anatomy abnormalities, like ventricular hypertrophy (Kligfield et al., 2007). The PhysioNet/Computing in Cardiology Challenge 2020 released a comprehensive 12-lead ECG dataset by aggregating multiple databases from around the world (Alday et al., 2020). The sources of ECG data included in this dataset are as follows:

- **CPSC**: This dataset was collected and published by the China Physiological Signal Challenge 2018 (CPSC 2018). It comprises 6,877 recordings sampled at 500 Hz, with varying lengths and an average duration of 16 seconds.
- **CPSC-Extra**: This dataset supplements CPSC 2018 with 3,450 additional recordings. It also has a sampling rate of 500 Hz and an average duration of 16 seconds but covers more types of cardiovascular diagnostics than CPSC.
- **St.Petersburg**: The third source is the public dataset from the St. Petersburg Institute of Cardiological Technics 12-lead Arrhythmia Database, Russia. It consists of 74 recordings collected at a high sampling rate of 257 Hz, with the longest average duration of 1,800

[4] https://timeseriesclassification.com/description.php?Dataset=ECG5000

seconds.

- **PTB**: The fourth source is the Physikalisch-Technische Bundesanstalt (PTB) Database in Brunswick, Germany. It includes 490 recordings collected at a high sampling rate of 1,000 Hz, with an average duration of 110.8 seconds.
- **PTB-XL**: This data was also collected by Physikalisch-Technische Bundesanstalt, but the sampling rate is 500 Hz. It contains 21,837 recordings, each with a length extract of 10 seconds.
- **G12ECG**: The last source is the Georgia 12-lead ECG Challenge (G12EC) Database from Emory University, Atlanta, Georgia, USA. It comprises 10,344 recordings, representing a large population from the Southeastern United States.

The aggregated data encompasses a total of 27 common diagnoses that are of clinical interest and are more likely to be recognisable from ECG recordings. Additionally, it is essential to highlight that each recording may be associated with multiple abnormalities, making this a multi-label classification task (Tsoumakas and Katakis, 2007; Yang et al., 2020). For the decentralised nature of this dataset, *i.e.*, it consists of six subsets collected from different areas, we leverage this data to validate the superiority of our proposed federated learning method, as presented in Chapter 7.3.

2.2.3 Dermoscopic images for skin lesion detection

With the continuous improvement in smartphone camera resolution, dermoscopy, a widely used tool in the field of dermatology, has evolved into mobile dermoscopy, a viable imaging method for dermatological practices (Kittler et al., 2002). Mobile dermoscopy involves using a dermoscope attached to a smartphone or tablet. This design offers enhanced portability while leveraging the advanced imaging capabilities of smartphones in conjunction with the magnification power provided by dermoscopy. For instance, mobile dermoscopes like MoleScope II are compatible with a wide range of smartphones and tablets, including but not limited to iPhone, iPad, and various Android models (Plüddemann et al., 2011).

Despite skin images being generally considered physiological time series data, they represent a crucial component of mobile health data within the field of dermatology and play a significant role in comprehending and assessing skin conditions. The International Skin Imaging Collaboration (ISIC)[5] is an academia and industry partnership designed to use digital skin imaging to help reduce skin cancer mortality. ISIC works to achieve its goals through the development and promotion of standards for digital skin imaging, and through engaging the dermatology and

[5] https://challenge.isic-archive.com/data/

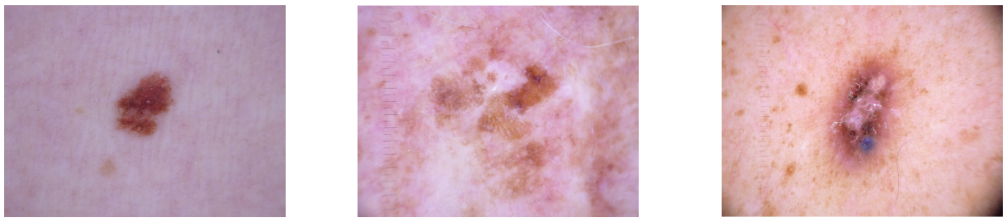


Figure 2.6: **Examples of HAM10000 skin image data.** The three displayed pathologies are visually distinguishable.

computer vision communities toward improved diagnostics. Some dermoscopic images with associated skin conditions are illustrated in Figure 2.6.

HAM10000 skin image data. Melanoma is the deadliest form of skin cancer. Among the precious challenges, HAM10000[6] is the dataset containing 10,015 dermoscopic skin tumour images taken from multiple devices and demographics (Tschandl et al., 2018). The image size is 600×450. The skin condition is labelled as one of the following classes: melanocytic nevi (67.1%), melanoma (11.1%), benign keratosis-like lesion (11.0%), basal cell carcinoma (5.1%), actinic keratoses (3.3%), vascular lesion (1.4%), or dermatofibroma (1.1%). For its severe imbalance character, we leverage this data to evaluate our work in quantifying the uncertainty with class imbalanced health data, as presented in Chapter 5.3.

In addition to the aforementioned physiological datasets, this thesis also includes a commonly used machine learning benchmark data for experiments. This benchmark data is the CIFAR10 image dataset (Krizhevsky et al., 2009). It is known for its large-scale nature, enabling us to manipulate it manually and simulate various data distributions for evaluation purposes.

CIFAR10 image data. CIFAR10 is widely used in the field of computer vision and machine learning. It stands for the Canadian Institute for Advanced Research 10-class dataset. CIFAR10 consists of 60,000 colour images, each measuring 32×32 pixels, and is divided into 10 classes. Each class contains 6,000 images. The dataset is split into a training set of 50,000 images and a test set of 10,000 images. The 10 classes in CIFAR10 represent different objects or animals, including airplanes, automobiles, birds, cats, deer, dogs, frogs, horses, ships, and trucks. In our studies, we down-sample this data, meaning we reduce the number of instances, to simulate various class distributions (see Chapter 5.3) and divide it into several subsets to mimic isolated data silos (see Chapter 7.3). The use of this benchmark data enables us to validate the generalisation of our proposed method in the broader field of machine learning.

[6] https://api.isic-archive.com/collections/212/

Chapter 3

Background and literature review

Machine intelligence is the last invention that humanity will ever need to make.

- Prof. Nick Bostrom

Director of Future of Humanity Institute, University of Oxford

3.1 Machine learning to model health from physiological data

To develop machine learning models for health diagnostics, pre-processed physiological data and health conditions are fed into a selected model to fit the optimal model parameters. This process is generally referred to as *model training*. Once the model is trained, it can be deployed to diagnose future physiological data, a process known as *model inference*. In the following section, we present the basics of developing a health diagnostic model.

3.1.1 Problem formulation

Formally, we refer to the physiological signals accompanied by health condition labels that are used to train the model as the *training set*. Sometimes, a fraction of the training set will be held out to identify the training-related hyper-parameters. This set is referred to as *validation set*[1]. The data that has not been employed to adjust model parameters is referred to as the *testing set*, which can be leveraged for evaluating model performance. To facilitate clarity in this thesis's

[1]Note that we adhere to the commonly accepted use of the term 'validation set' from the machine learning literature. In contrast, in a medical context, a 'validation set' often refers to a subset of data or cases used to assess the reliability and accuracy of a diagnostic test, treatment plan, or predictive model. This differs from its application in this thesis.

presentation, we consistently employ the following notations to denote both the data and the model:

Problem formulation and notations. Assuming the availability of a training dataset comprising pre-processed physiological data denoted as $\mathcal{D} = \{(x^{(i)}, y^{(i)})\}_{i=1}^{N}$. Here, $x^{(i)}$ represents the input physiological sample, $y^{(i)}$ corresponds to the health condition, *i.e.*, the label of the sample, among a total of C categories and N training samples. The number of training samples for the c-th class is termed as N_c in this thesis. Due to the difficulty of annotations and inherent class imbalance nature in health-related datasets (refer to Challenge (*i*) in Chapter 1.2), the training set is usually limited and tends to be skewed, resulting in varying values of N_c across classes. In this context, a class is referred to as a majority class if $N_c \gg N/C$, while it is considered a minority class otherwise. Both binary and multi-class diagnostics are considered. For *binary diagnostics*, $C = 2$ with one class for healthy controls and the other for the unhealthy group. For *multi-class diagnostics*, $C > 2$ with each class associated with a specific health condition. $\mathcal{D}$ will be employed to train a model parameterised by θ that can predict $y^{(i)}$ for any given sample $x^{(i)}$ from a testing set. Also because of privacy concerns, the training dataset $\mathcal{D}$ could be distributed in K places, *i.e.*, $\mathcal{D}_1, ..., \mathcal{D}_K$, rather than being centrally available (as specified by Challenge (*iii*) in Chapter 1.2). Additionally, the model will provide a predictive confidence $u^{(i)}$ associated with the prediction. The confidence measurement should reflect how reliable the diagnostic result is for each input (*i.e.*, addressing Challenge (*ii*) in Chapter 1.2).

In this thesis, particular emphasis is placed on deep learning due to its outstanding performance in the literature. Deep learning primarily entails the utilisation of deep neural networks for data modelling. Figure 3.1 offers an overview of the classical training process for a deep neural network designed for disease diagnostics based on physiological data. The following section provides a comprehensive explanation of each phase of this procedure.

3.1.2 Foundations of deep neural networks

There are several basic concepts in deep neural networks (DNNs), from model architectures to model optimisation, that lay the foundation for deep learning applications. We introduce them in the following sections.

Deep neural networks

The initial step in developing a deep learning-driven health diagnostic model is to choose an appropriate model architecture based on the characteristics of the physiological data. The history of neural networks can be traced back to the 1950s, with the invention of the perceptron (Rosenblatt, 1958), which laid the foundation for today's modern Deep Neural Networks (DNNs) (LeCun et al., 2015; Deng et al., 2014). In the present day, widely adopted deep neural network

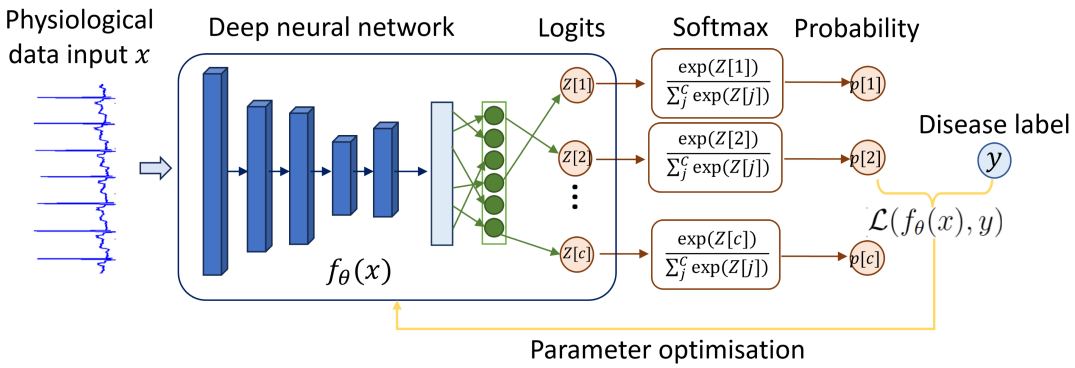


Figure 3.1: **An illustration of model development**. A physiological data sample is input into a deep neural network, which derives logits for all classes. Through the *Softmax* operation, the model produces corresponding probabilities. Based on the predicted probability and the ground-truth label, the loss is calculated and then back-propagated to optimise the model's parameters.

architectures include:

- *Convolutional Neural Networks* (CNNs): CNNs are primarily used for image and video-related tasks. They consist of convolutional layers that extract local features from input data, followed by pooling layers for spatial down-sampling. The crucial component within the convolutional layer is the convolutional kernel, which is a small trainable matrix. The convolution operation involves sliding the kernel over the input data and computing the element-wise dot product at each position. This operation results in a feature map that highlights specific patterns or features in the input data. The kernel is typically two-dimensional for image-format data, but it can also be one-dimensional, allowing it to be used for time series data to extract features. (O'Shea and Nash, 2015; Gu et al., 2018). In this thesis, we employ the two-dimensional CNN to model the spectrograms of physiological audio (refer to Chapter 4.3, and 6.3), and the one-dimensional CNN to model the ECG signals (refer to Chapter 5.3 and 7.3).

- *Recurrent Neural Networks* (RNNs): RNNs are suitable for time series data processing, such as natural language processing and speech recognition. They have recurrent connections that allow information to persist over time, making them capable of handling sequential dependencies (Medsker and Jain, 2001). *Long Short-Term Memory* (LSTM) (Graves and Graves, 2012) and *Gated Recurrent Unit* (GRU) (Chung et al., 2014) are popular variations of RNNs that address the vanishing gradient problem and improve information retention. GRUs use a gating mechanism to control the flow of information through the network for each time step, and also maintain a memory cell that captures information from previous time steps.

- *Transformers*: Transformers have gained significant attention for natural language pro-

cessing tasks. They employ self-attention mechanisms to capture contextual relationships between words or tokens in a sequence (Vaswani et al., 2017). Transformers have achieved state-of-the-art performance in tasks such as machine translation, text generation, and sentiment analysis. The Transformer architecture is commonly used in models like BERT (*Bidirectional Encoder Representations from Transformers*) (Sun et al., 2019) and GPT (*Generative Pre-trained Transformer*) (Luo et al., 2022).

In summary, the above are a few examples of typical deep neural network architectures, and the choice of architecture depends on the specific task, dataset, and domain. Researchers continue to explore and develop new architectures to address various challenges and improve performance in different domains.

In deep neural networks, besides the aforementioned fully connected layers, convolutional layers, and self-attention layers, non-parametric activation functions play a crucial role. An activation function is a mathematical operation applied to a node (or 'neuron') in a neural network, transforming the input signal into an output signal for that node. It is a vital component in neural networks, determining whether a neuron should be activated or not based on the weighted sum of its inputs.

Activation functions are important because of: *i) Non-linearity*. Activation functions introduce non-linear properties to the network. Without non-linearity, a neural network, regardless of the number of layers it has, would behave just like a single-layer network because linear operations are closed under composition. Non-linear functions enable neural networks to learn complex mappings from inputs to outputs, allowing them to perform tasks such as image recognition, language translation, and playing complex games. *ii) Control of activation.* They determine whether a neuron should be activated by calculating the weighted sum and further adding bias to it. They are used to introduce non-linearity into the model so that it can learn more complex decision boundaries.

There are several types of activation functions used in deep learning, each with its own characteristics and applications. The two commonly used ones for classification tasks are *ReLU* (Rectified Linear Unit) and *Softmax*, which are introduced below.

ReLU function

It is one of the most widely used activation functions in deep learning models, especially in CNNs. The function is defined as follows:

$$\text{ReLU}(x) = \max(0, x) \tag{3.1}$$

This means that for each input x, the *ReLU* function outputs x if x is greater than zero, and outputs zero otherwise. Graphically, the *ReLU* function is a straight line that passes through the origin $(0, 0)$ with a slope of 1 for all positive values of x, and a slope of 0 for all negative values of x.

ReLU helps in mitigating the vanishing gradient problem, which is a situation where the gradients become too small for the network to learn effectively. Since the gradient for positive inputs is always 1, this ensures that the network continues to learn as long as there are positive inputs.

Softmax function

As depicted in Figure 3.1, DNNs produce a set of raw scores, often referred to as *logits*. On the output side, the *Softmax* function is typically employed to transform these *logits* into a probability distribution across multiple classes. This mathematical function has trainable parameters but plays a crucial role in classification tasks.

Specifically, for health diagnostics, the deep neural network generates *logits* for each disease class based on the input physiological data. These *logits* represent the non-normalised scores assigned to each class, and they might not be directly interpretable as probabilities. The *Softmax* function transforms these raw scores into a probability distribution by exponentiating the scores and normalising them. Mathematically, for a given input $x^{(i)}$, the logit vector $\boldsymbol{z}^{(i)}$ is transferred into the categorical probability vector $\boldsymbol{p}^{(i)}$ by,

$$\begin{aligned} \boldsymbol{z}^{(i)} &= f_\theta(x^{(i)}), \\ \boldsymbol{p}^{(i)}[c] &= \frac{\exp(\boldsymbol{z}^{(i)}[c])}{\sum_{j=1}^{C} \exp(\boldsymbol{z}^{(i)}[j])}, \end{aligned} \tag{3.2}$$

where $[c]$ presets the score for class c, f_θ denotes the model function parameterised by θ, and $\sum_{c=1}^{C} p_c^{(i)} = 1$. Correspondingly, the final prediction $\hat{y}^{(i)}$ is the class with the maximum probability,

$$\hat{y}^{(i)} = \arg\max_c \ \boldsymbol{p}^{(i)}[c]. \tag{3.3}$$

Parameter optimisation

As show in Figure 3.1, to train the neural network and learn its parameters θ from the training set, a differentiable loss function $J(\theta)$ is introduced as the objective function for an optimisation algorithm. The parameters which minimise $J(\theta)$, are considered to best fit the model. A very common guiding principle for training is to reduce the *empirical error*, which refers to the difference between the predicted output of a model and the actual target output for a set of training data points (Jordan and Mitchell, 2015).

For health diagnostics, which is fundamentally a classification problem, the commonly used loss function is the *cross-entropy loss* (Rubinstein and Kroese, 2004). It measures the dissimilarity between predicted class probabilities and the true one-hot encoded class labels[2]. Specifically, $J(\theta)$ can be written as an average over the training set as follows,

$$J(\theta) = \mathbb{E}_{(x,y)\sim\mathcal{D}}[\mathcal{L}(f_\theta(x), y)], \tag{3.4}$$

where $\mathcal{D}$ is training set that presents the empirical data distribution, and $\mathcal{L}(f_\theta(x), y) = -\log \boldsymbol{p}[y]$ is the per-example cross-entropy value ($\boldsymbol{p}[y]$ denotes the probability for the ground-truth class in Eq. (3.2)).

To minimise $J(\theta)$, gradient descent is employed as a fundamental optimisation algorithm by iteratively adjusting the parameters in the direction of $J(\theta)$'s gradient, termed by $\nabla_\theta J(\theta)$, that leads to a minimum of the function. There are variations of gradient descent, but the most common implementation is Stochastic Gradient Descent (SGD). In SGD, the expectation of the gradient (refer to Eq. (3.4)) is computed by randomly sampling a small number of examples from the training set, *i.e.*, a batch $\mathcal{B}$, then taking the average over only those examples. The updating the parameters over one batch is formulated as,

$$\theta = \theta - \lambda \frac{1}{|\mathcal{B}|} \nabla_\theta \sum_{(x,y)\sim\mathcal{B}} \mathcal{L}(f_\theta(x), y), \tag{3.5}$$

where λ is a coefficient that controls the learning rate of the parameters. Since SGD only uses a subset of the data in each iteration, the algorithm can make updates more frequently, which can help escape local minima and converge faster. $\nabla_\theta J(\theta)$ is calculated using *back-propagation* by applying the chain rule for partial derivatives, starting from the output layer towards the input (Jordan and Mitchell, 2015; Hecht-Nielsen, 1992).

Adam is an another very popular optimiser in deep learning (Kingma and Ba, 2014). Unlike SDG using a fixed learning rate λ, Adam adjusts the learning rate for each parameter individually based on estimates of the first (mean) and second (uncentered variance) moments of the gradients. This allows it to handle sparse gradients on noisy problems and is more efficient compared to SDG.

Beyond the typical optimisation method known as fully supervised learning, there are many other learning paradigms such as transfer learning, semi-supervised learning, and self-supervised learning, which prove particularly useful when the labelled training set is limited. Transfer learning is a technique where a model developed for a specific task is repurposed as the starting point

[2] An one-hot encoded class label is a C-dimensional vector. For a given data point, the corresponding one-hot encoded vector is all zeros except for the index that corresponds to the class label, which is set to 1. For example, suppose we have three classes: A, B, and C. The one-hot encoding for each class would look like this: Class A: [1, 0, 0]; Class B: [0, 1, 0]; Class C: [0, 0, 1].

for a model on a second task, thus enhancing performance and reducing the training time and resources required (Weiss et al., 2016; Zhuang et al., 2020). Semi-supervised learning utilises a small amount of labelled data in conjunction with a large amount of unlabelled data during the training process (Van Engelen and Hoos, 2020; Yang et al., 2022b). This approach is especially beneficial when acquiring a comprehensive set of labelled data is costly or time-consuming, but unlabelled data is plentiful. Self-supervised learning, on the other hand, generates its own supervision from the input data, making full use of the data to learn valuable representations or features directly from the data itself (Liu et al., 2021b). These techniques are also widely used in physiological data modelling (Eldele, 2023).

In addition to learning the model parameters, training a deep learning model also involves making decisions about the values of hyper-parameters, such as the depth of the network or the size of the architecture. Usually, a small proportion of the training set (as the mentioned *validation set*) will be used to determine the optimal hyper-parameters. Specifically, a grid search is performed over a range of hyper-parameters, during which the model is trained with various hyper-parameter settings, and its performance on the validation set is recorded. Finally, the hyper-parameter configuration that yields the best performance is selected, and the corresponding model is evaluated on the testing set.

3.2 Machine learning-driven health diagnostics using physiological data

Chapter 2 introduced the three physiological data modalities analysed in this thesis: respiratory audio, ECG, and dermoscopic images. The following section presents an overview of the machine learning techniques designed for these data types for health diagnostics purposes.

3.2.1 Acoustic machine learning for respiratory health

For centuries, medical professionals have utilised audio as a diagnostic technique in assessing the respiratory system (Mascolo, 2020; Bohadana et al., 2014; Hanna and Silverman, 2002), primarily through the use of stethoscopes. Acquiring auscultation skills often entails several years of training for medical practitioners. However, the widespread availability of mobile devices equipped with high-fidelity built-in microphones has recently presented an unprecedented opportunity to collect acoustic signals from the population. This development has sparked our enthusiasm for leveraging the power of AI to automatically analyse this audio data, with the aim of respiratory health screening (Hadjitodorov and Mitev, 2002; Mukherjee et al., 2021; Srivastava et al., 2021; Pramono et al., 2016; Hao et al., 2013; Schuller, 2013; Yu and Li, 2017). Existing studies have primarily developed two types of methods: hand-crafted feature-based

models and end-to-end deep learning models.

In feature-based models, temporal features, particularly prosodic features like pitch, duration, intensity, harmonics-to-noise ratio, jitter, and shimmer, are extensively used for detecting abnormal sounds (Hadjitodorov and Mitev, 2002). Additionally, spectral features derived from the spectrogram have been created and have shown promising performance across various related applications (Mukherjee et al., 2021; Srivastava et al., 2021; Pramono et al., 2016; Hao et al., 2013). These features serve as inputs for subsequent classifiers in the diagnostic process. Feature-based models tend to be interpretable, but their effectiveness is limited by the expertise and knowledge embedded in their development.

End-to-end deep learning methods use audio waves (Sharan, 2023) or their corresponding spectrograms (Ren et al., 2020) as inputs for classification. These models consist of multiple layers that can automatically capture the complex relationship between the input and the output labels (Schuller, 2013; Yu and Li, 2017). While lacking in interpretability, these models usually outperform the aforementioned feature-based methods. For example, Shi *et al.* designed CNN models to classify various lung sounds, including wheeze, squawk, stridor, and crackle, achieving an accuracy of over 95% (Shi et al., 2019; Bardou et al., 2018). Altan *et al.* proposed a deep belief network derived from the Hilbert Transform via multi-channel lung sounds to diagnose COPD (chronic obstructive pulmonary disease), with a sensitivity of 91% and a specificity of 96.33% (Altan et al., 2019).

These existing studies have paved the way for the development of audio-based respiratory health screening models. While existing work focuses on identifying useful acoustic features, our research pays more attention to the problems of class imbalance, model overconfidence, and data privacy, which remain unexplored.

3.2.2 Classification of electrocardiogram signals

The classification of ECG signals is crucial in the clinical diagnosis of heart disease. A significant challenge in diagnosing heart disease using ECG is the variability among individuals; a normal ECG can differ from person to person, and a single disease may not exhibit consistent signs across different patients' ECG signals. Additionally, two distinct diseases may present similar effects on normal ECG signals. These issues complicate the diagnosis of heart disease. Consequently, employing pattern classification techniques can enhance the diagnosis of ECG arrhythmias in new patients (Jambukia et al., 2015).

Similar to the acoustic machine learning approaches discussed in the previous section, the classification of ECG signals can also be categorised into traditional feature-based and end-to-end deep learning-based approaches. As depicted in Figure 2.5(a), an ECG signal comprises several

beats, and each beat contains a P wave, QRS complex, and T wave. Each peak (P, Q, R, S, T, and U), interval (PR, RR, QRS, ST, and QT), and segment (PR and ST) of the ECG signals possesses standard amplitude or duration values. These peaks, intervals, and segments are referred to as ECG features (Mar et al., 2011). These features act as inputs for classifiers in the diagnostic process. The literature indicates that these features can achieve satisfactory accuracy in detecting many heart diseases, such as Atrial Fibrillation (a common and serious condition characterised by an irregular and often rapid heart rate that can lead to blood clots in the heart), when the quality of ECG signals is high (Orphanidou et al., 2015).

While the feature-based model provides better explainability, end-to-end deep learning models tend to achieve better generalisation and are more robust to signal noise (Jambukia et al., 2015). One-dimensional CNNs are the most commonly used architecture for ECG signal classification (Hygrell et al., 2023; Attia et al., 2019). Although other architectures, such as RNNs, are also applicable, they do not show a significant advantage for ECG classification. This is primarily due to the efficient learning capabilities of CNNs with limited training data (Xiong et al., 2017).

Although numerous studies have demonstrated the effectiveness of machine learning and deep learning in detecting cardiovascular diseases using ECG data, most of these studies have relied on well-curated, centralised datasets. This approach often overlooks crucial aspects such as the model's reliability in the wild and data privacy protection issues. This thesis aims to address these gaps.

3.2.3 Classification of dermoscopic images

Classifying dermoscopic images using machine learning and deep learning techniques has become a significant area of research in dermatology, aiming to improve the diagnosis and screening of skin diseases, including skin cancer (Grignaffini et al., 2022).

Early efforts in classifying dermoscopic images relied on traditional machine learning algorithms. These approaches typically involve handcrafted feature extraction, where specific characteristics like colour, texture, shape, and border are identified and used as inputs for classifiers (Barata et al., 2018; Talavera-Martinez et al., 2019). Commonly used machine learning classifiers include Support Vector Machines (SVM), k-Nearest Neighbors (k-NN), and Random Forests. While effective to a degree, the performance of these methods heavily depends on the quality and relevance of the extracted features (Javed et al., 2020).

The advent of deep learning has dramatically enhanced the ability to classify dermoscopic images automatically. Deep learning models, particularly CNNs, have shown exceptional performance in image recognition tasks, including dermoscopic image classification (Iqbal et al.,

2021). Unlike traditional machine learning, CNNs automatically learn hierarchical feature representations from the images, eliminating the need for manual feature extraction. Several CNN architectures have been explored for dermoscopic image classification, including AlexNet (Pomponiu et al., 2016; Iyatomi et al., 2011), VGGNet (Alfed et al., 2015), ResNet (Mikołajczyk et al., 2017), and Inception models (Saez et al., 2014). These models have been pre-trained on large general image datasets and fine-tuned on dermoscopic images to achieve high levels of accuracy in identifying various skin conditions, such as melanoma, basal cell carcinoma, and benign lesions (Lopez et al., 2017).

Classifying dermoscopic images through machine learning and deep learning presents a promising avenue for automating the diagnosis of skin diseases. Current research efforts are directed not only towards enhancing the accuracy and efficiency of these classification models but also towards overcoming practical challenges associated with model reliability. Our study specifically explores the effects of class imbalance and model overconfidence, offering effective solutions to these issues.

3.3 Advanced deep learning paradigms

In the preceding section, we presented the typical machine learning and deep learning methods for respiratory audio, ECG signals, and skin images. However, as detailed in Chapter 1.2, training robust and high-performing deep neural networks for healthcare faces numerous challenges arising from the complexities of data collection and the safety-critical nature of diagnostic applications. This section introduces several advanced training paradigms that lay the groundwork for our proposed solutions aimed at addressing these challenges.

3.3.1 Long-tailed learning for class imbalanced data

The first challenge we identified in Chapter 1.2 is the prevalence of insufficient and imbalanced physiological data for machine learning research. Such a character can pose challenges for several reasons: *(i) biased model performance.* Deep learning models are trained to optimise a loss function by minimising errors. In the case of class imbalance, a model can achieve high accuracy by simply predicting the majority class most of the time, even though it fails to capture the patterns and characteristics of minority classes. As a result, the model's performance may be biased towards the majority class, leading to poor predictions for the minority class; *and (ii) insufficient learning from minority classes*. Deep learning models typically require a sufficient number of samples to learn robust representations and patterns. With class imbalance, the limited number of samples from minority classes can hinder the model's ability to adequately learn their distinctive features. Consequently, the model may struggle to generalise well to unseen data or make accurate predictions for minority class instances.

In healthcare applications, it is crucial to accurately detect infrequent yet significant health events within predominantly healthy data (Mazurowski et al., 2008; Saini and Susan, 2019; Afzal et al., 2019). Therefore, addressing and mitigating the adverse effects of class imbalance is essential to ensure reliable and effective health-related analyses and predictions.

Long-tailed learning is the main technique that allows training models from the class imbalanced data (Zhang et al., 2023c), which covers two categories: *Data-level* methods and *algorithm-level* methods. The simplest *data-level* methods are random under-sampling (RUS) which discards samples from the majority classes and random over-sampling (ROS) which re-samples from the minority classes during training (Van Hulse et al., 2007). For example, the up-sampling-based data augmentation method SMOTE (synthetic minority over-sampling technique) has been widely adopted for the minority classes in health applications (Chawla et al., 2002; Rahman and Davis, 2013; Han et al., 2021a). Those are infeasible when the data imbalance is extreme (Johnson and Khoshgoftaar, 2019). Synthetic generation (He et al., 2008; Jacsó, 2005) or interpolation (Chawla et al., 2002) to increase the minority samples are also explored. However, they are sensitive to imperfections in the generated data and hard to generalise. *Algorithm-level* methods modify the training procedure by introducing cost-sensitive losses or scaling the classification thresholds. Well-known implementations include Class-balanced loss (Cui et al., 2019), and focal loss (Lin et al., 2017). This type of method usually involves hyper-parameters that need to be searched during training.

Relating to our work. Addressing class imbalance is crucial in health applications, and it is observed that class imbalance frequently coexists with other challenges in physiological data, including model overconfidence and data privacy. This thesis endeavours to comprehensively tackle these challenges within unified frameworks. Chapters 4 and 5 delve into methods designed to calibrate deep learning training when confronted with imbalanced data. In Chapter 6, we introduce methodologies specifically tailored for handling decentralised imbalanced physiological data.

3.3.2 Uncertainty quantification for model calibration

Another challenge, as identified in Chapter 1.2, pertains to model overconfidence, a situation where the model yields unreliable confidence to the extent of its unknowns. In a prior study, it has been empirically demonstrated that deep neural networks although achieving better categorical predictions, are poorly calibrated compared to shallow neural networks (Guo et al., 2017). This is visualised in Figure 3.2: The top row shows the distribution of prediction confidence (*i.e.*, probabilities associated with the predicted label) as histograms. The average confidence of LeNet closely matches its accuracy, while the average confidence of the ResNet is substantially higher than its accuracy (it is so-called *model overconfidence*). This is further illustrated in the

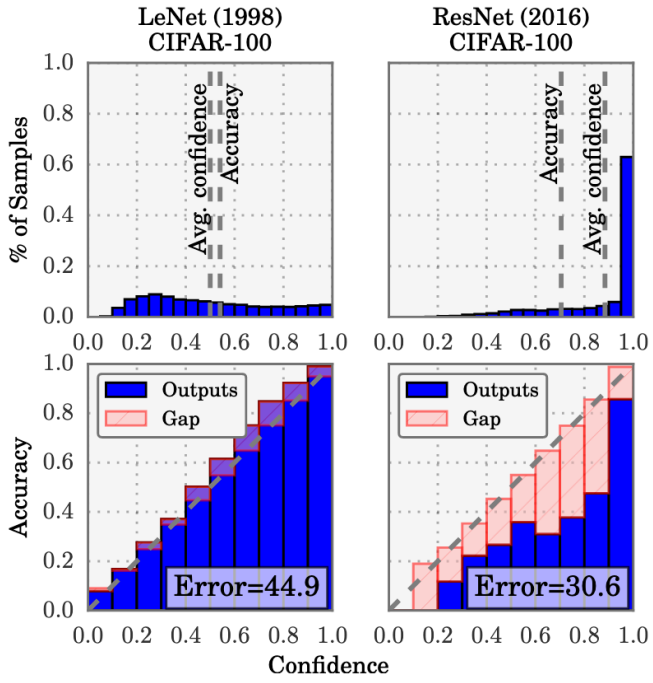
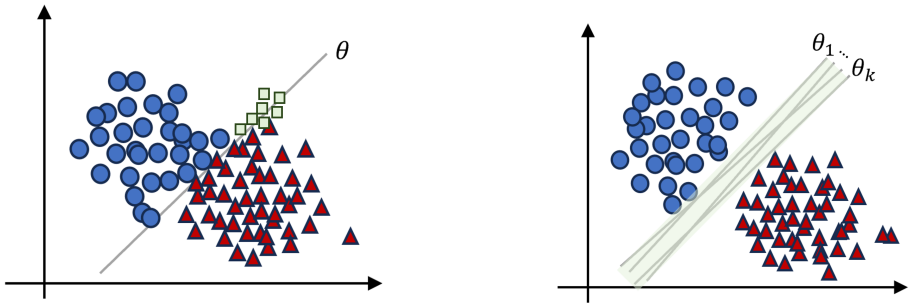


Figure 3.2: **A comparison of confidence histograms (top) and reliability diagrams (bottom) between shallow and deep neural networks.** The task involves image classification using the CIFAR100 dataset. LeNet (left) is a 5-layer CNN model (LeCun et al., 1998) and ResNet (right) is a 110-layer CNN model (He et al., 2016).

bottom row reliability diagrams (Niculescu-Mizil and Caruana, 2005), which show accuracy as a function of confidence. It can be observed that LeNet is well-calibrated, as confidence closely approximates the expected accuracy (*i.e.*, the bars align roughly along the diagonal). On the other hand, the ResNet's accuracy is better but does not match its confidence.

A poorly calibrated model can produce incorrect predictions with unwarranted overconfidence, which is particularly concerning in the context of health diagnostics. To address this problem, uncertainty estimation serves the crucial purpose of quantitatively measuring the reliability of a model's predictions. It plays a pivotal role in facilitating the deployment of deep learning for real-world healthcare applications (Abdar et al., 2021).

There are mainly two types of uncertainty in deep learning: The uncertainty that encompasses the noise inherited from data is referred to *aleatoric uncertainty*, and the uncertainty due to the insufficient knowledge of a model is known as *epistemic uncertainty* (Gawlikowski et al., 2021). An illustration of the two types of uncertainty is given in Figure 3.3. Unlike *aleatoric uncertainty* (also referred to as *data uncertainty*), which is inherent and irreducible, *epistemic uncertainty* (as termed as *model uncertainty*) can be reduced by gathering more data, improving data quality, or developing more sophisticated models that better capture the underlying processes. The overall uncertainty encompassing these two types is known as *predictive uncertainty*. Many



(a) Aleatoric uncertainty. (b) Epistemic uncertainty.

Figure 3.3: **An illustration of uncertainty**. Consider the development of a machine learning model aimed at categorising data as example. The presence of noise, perturbations, and biases within the data introduces a layer of data uncertainty, stemming from inherent randomness and noise, which complicates the task of making precise predictions. This phenomenon is depicted through the use of green markers in panel (a). Additionally, model uncertainty arises from either insufficient knowledge regarding the optimal model or the absence of sufficient training data. This aspect is symbolised by the shadow observed in panel (b).

uncertainty-driven applications like misclassification identification and out-of-distribution detection require the model to capture both *aleatoric uncertainty* and *epistemic uncertainty* (Shen et al., 2023).

In the following section, we first examine the limitations of the commonly used *Softmax*-based neural network for uncertainty quantification, and then we introduce four advanced approaches for more accurate uncertainty estimation. These methodologies are further explored in Chapters 4 and 5.

Standard neural network with *Softmax*

Chapter 3.1 introduces a standard way of building deep neural networks with *Softmax* function for classification probabilities. Yet, the parameters are *deterministic*, and thus this approach is limited in providing the variance of the prediction. Therefore, such probability cannot reflect epistemic uncertainty as shown in Figure 3.3(b).

Furthermore, the *Softmax* function is known for its tendency to overestimate probabilities. In Eq. (3.2), the *Softmax* operation involves exponentiating the logits and normalising them into a probability distribution, but this process often results in over-estimations, especially for unseen classes. As a consequence, this can lead to unreliable estimations of aleatoric uncertainty.

Numerous studies, including those by (Gal and Ghahramani, 2015; Louizos and Welling, 2017), as well as the research presented in this thesis, have shown that neural networks tend to provide

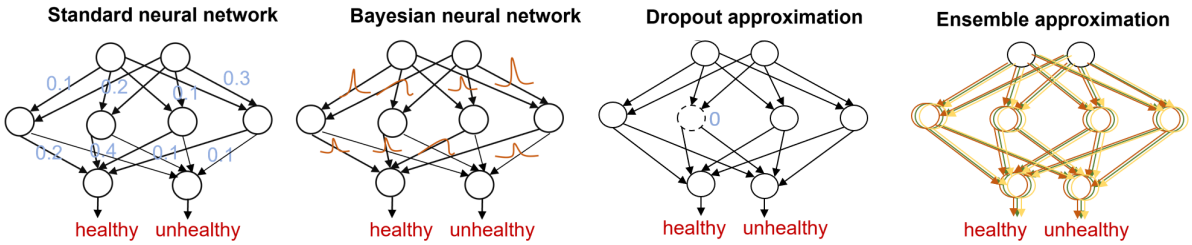


Figure 3.4: **An illustration of uncertainty quantification models**. Standard neural networks and Bayesian neural networks with their approximations are compared.

overconfident classifications and probabilities for inputs that fall outside the training distribution. Additionally, even within the training distribution, misclassified samples often yield high classification probabilities when computed using the *Softmax* function. This suggests that *Softmax* is insufficient for effectively incorporating uncertainty into predictions.

Bayesian neural networks

Bayesian inference allows the learning of a probability distribution over possible neural networks, and such a type of neural network is known as a Bayesian neural network (Blundell et al., 2015). Compared to standard deterministic neural networks, Bayesian methods learn a posterior distribution of the model parameters based on a learnable prior distribution that is estimated from the observed data (see Figure 3.4). Specifically, given the observed data set $D = \{(x, y)\}$, a conditional likelihood $p(y|x, \theta)$ tells how each model θ explains the relation of input x and target y. With the posterior distribution $p(\theta|D)$, for a new test data point x', the predictive distribution can be derived as,

$$p(y|x') = \int p(y|x', \theta)p(\theta|D)d\theta, \tag{3.6}$$

$$p(\theta|D) = \frac{p(y|x, \theta)p(\theta)}{\int p(y|x, \theta)p(\theta)d\theta}. \tag{3.7}$$

$p(y|x')$ captures both epistemic and aleatory uncertainty.However, deep neural networks typically involve millions of parameters, rendering it impractical to compute the exact posterior distributions, such as $p(\theta|D)$ and $p(y|x')$. Consequently, various approaches for posterior distribution approximation have been proposed, including Markov chain Monte Carlo (MCMC) (Neal, 2012; Chen et al., 2014; Welling and Teh, 2011) and variational inference (Blei et al., 2017; Graves, 2011; Lopez et al., 2018). Despite their utility, these methods often face challenges of slow and inefficient posterior estimation.

Monte Carlo Dropout (MCDropout)

In modern deep learning, Dropout is widely used to mitigate overfitting (Baldi and Sadowski, 2013; Srivastava et al., 2014). However, more recently, Gal *et al.* introduced a new perspective by leveraging variational distributions to interpret dropout in the forward phase as approximate Bayesian inference (Gal and Ghahramani, 2016a,b). Typically, Dropout is applied during training to preserve the model's capacity. Nevertheless, if Dropout is kept active during inference, the predictive probability can be calibrated using randomly sampled model weights, making it an estimation of uncertainty. This approach significantly reduces the computational cost during training.

Take the fully connected neural network layer l for an example. The normal operation is,

$$x^{(l+1)} = \sigma(x^{(l)}W^{(l)} + b^{(l)}), \tag{3.8}$$

where $x^{(l)}$ and $x^{(l+1)}$ are the input and output of layer l parameterised by weight matrix $W^{(l)}$ and bias vector $b^{(l)}$, and σ is the non-linear activation function. With Dropout, the operation is instead,

$$\begin{aligned} z^{(l)} &\sim Bernoulli(\cdot|p^{(l)}), \\ \tilde{W}^{(l)} &= diag(z^{(l)})W^{(l)}, \\ x^{(l+1)} &= \sigma(x^{(l)}\tilde{W}^{(l)} + b^{(l)}). \end{aligned} \tag{3.9}$$

Using dropout at the layer l is mathematically equivalent to setting the rows of the weight matrix $W^{(l)}$ for that layer to zero. This is controlled by variable $z^{(l)}$ from the Bernoulli distributed random variables with some probabilities $p^{(l)}$. The $diag(\,\cdot\,)$ maps vectors to diagonal matrices. The above operation can be generalised to other types of layers as well (Gal and Ghahramani, 2016a,b).

Overall, the described dropout operations convert a deterministic neural network parameterised by θ into a random Bayesian neural network with random variables $\tilde{\theta}$, which equates to a neural network with a statistical model without using the Bayesian approach explicitly. Finally, to approximate the predictive distribution $p(y|x, \theta)$, Monte Carlo (MC) sampling of the random variables $\tilde{\theta}$ is performed,

$$p(y|x') = \frac{1}{T}\sum_{i=1}^{T} p(y|x', \tilde{\theta}^i), \tag{3.10}$$

where T is the number of MC samples. It is equivalent to performing T stochastic passes. This method is commonly known as variational dropout, and in this thesis, we refer to it as MCDropout.

Figure 3.4 provides an example of one forward pass, where the activation of an intermediary

layer is randomly set to zero. As such, MCDropout can be easily implemented in any neural network that has been trained with dropout, utilising Monte Carlo sampling and requiring multiple runs of the entire network. Despite its simplicity, several studies have criticised MCDropout as a sub-optimal Bayesian approximation (Osband, 2016; Hron et al., 2017). Furthermore, using a fixed dropout rate instead of optimising the variational parameter can result in an arbitrarily poor approximation. The actual performance of MCDropout may vary depending on the task difficulty, the quality of the training data, and the quantity of available data.

Ensemble learning

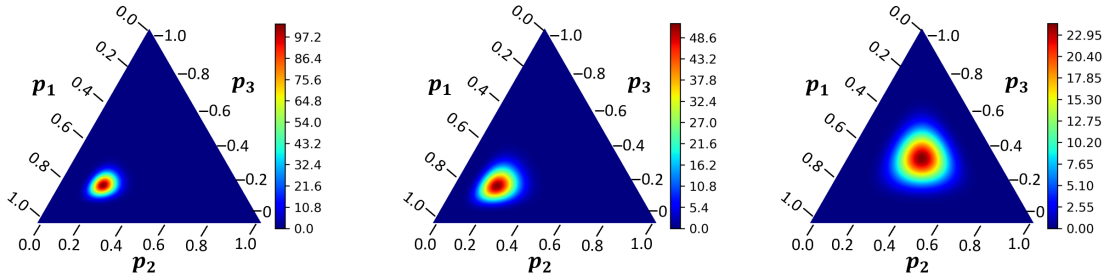
The ensemble, also known as a frequentist method, represents another type of approximation for Bayesian neural networks. Instead of learning a closed-form distribution for model parameters, ensemble approaches only require a limited number of models, which is computationally tractable (Ganaie et al., 2022). Herein, uncertainty stems from how the prediction is expected to change with different network structures or training data. An ensemble consists of multiple models with the same network structure but trained from different instances re-sampled from the dataset (Dietterich, 2000; Zhou, 2012). One of the popular strategies is bagging (known as bootstrapping), where ensemble members are trained on different bootstrap samples of the original training set. Ensemble methods can capture both aleatoric and epistemic uncertainty, with the predictive probability formulated by a uniformly weighted combination of the outputs from N_m models, denoted as,

$$p(y|x') = \frac{1}{N_m} \sum_{i=1}^{N_m} p(y|x', \theta^i). \tag{3.11}$$

Through extensive experiments on synthetic and real-world data, Lakshminarayanan *et al.* proved that a simple ensemble can produce well-calibrated uncertainty estimates which are as good or better than approximate Bayesian neural networks (Lakshminarayanan et al., 2017). In addition, with the estimated uncertainty, the model is able to detect out-of-training distribution for more reliable prediction (Lee et al., 2018). Overall, the ensemble is a comprehensive, robust, and accurate method for posterior inference, although obtaining a bootstrap ensemble of size N_m is computationally intense as N_m times as training a single model.

Evidential deep learning

Compared to deep ensembles and Bayesian neural networks (Gawlikowski et al., 2021), evidential deep learning (EDL), a recently emergent method, has demonstrated notable efficiency and effectiveness (Malinin and Gales, 2018; Sensoy et al., 2018; Charpentier et al., 2020; Kopetzki et al., 2021). EDL also offers the advantage of leveraging pre-trained models for uncertainty quantification, particularly in scenarios with limited data availability. The core principle of



(a) $\boldsymbol{\alpha}$ = [60,20,20] $\mathbb{E}[\boldsymbol{p}]$ = [0.6,0.2,0.2]

(b) $\boldsymbol{\alpha}$ = [30,10,10] $\mathbb{E}[\boldsymbol{p}]$ = [0.6,0.2,0.2]

(c) $\boldsymbol{\alpha}$ = [10,10,10] $\mathbb{E}[\boldsymbol{p}]$ = [0.33,0.33,0.33]

Figure 3.5: **Three-class Dirichlet distribution**. (a) and (b) point to the same predicted class, but (a) is sharper so it is more confident while (b) is more uncertain. (c) shows an example that is certain to none of the classes.

EDL is to learn the evidence for classification and establish a distribution over predictive probabilities. This allows for the quantification of the inherent uncertainty present in the output distribution. Typically, a Dirichlet distribution is employed as the posterior distribution, as it serves as the natural conjugate posterior for the categorical distribution. This characteristic has led to EDL being referred to as the Dirichlet-based uncertainty quantification method (Ulmer, 2021). The integration of EDL into existing health diagnostic models is straightforward, as it simply requires replacing the output layer of the model with a Dirichlet distribution output.

Formally, different from the *Softmax*-based model, EDL leverages Dirichlet distribution $\boldsymbol{q}^{(i)}$, *i.e.,* the distribution over the categorical probability $\boldsymbol{p}^{(i)}$, to achieve prediction and uncertainty quantification simultaneously (Hastie et al., 2009; Murphy, 2012). The Dirichlet distribution is used because it is the natural conjugate posterior of multinomial distribution (*i.e.*, the probability $\boldsymbol{p}^{(i)}$ can be regarded as a multinomial distribution). Underpinned by the Bayesian rule, EDL aims to capture the classification evidence $\boldsymbol{l}^{(i)}$ by the deep model and then transform a uniform prior Dir($\mathbf{1}$) into the posterior $\boldsymbol{q}^{(i)} = \text{Dir}(\boldsymbol{\alpha}^{(i)})$, with $\boldsymbol{\alpha}^{(i)} = \mathbf{1} + \boldsymbol{l}^{(i)}$ (Murphy, 2012). More specifically, the posterior $\boldsymbol{q}^{(i)} = \text{Dir}(\boldsymbol{\alpha}^{(i)})$ is parameterised by $\boldsymbol{\alpha}^{(i)} = [\alpha_1^{(i)}, \alpha_2^{(i)}, ..., \alpha_C^{(i)}]$ for C classes, where $\alpha_c^{(i)} = 1 + l_c^{(i)}$.

The posterior Dirichlet distribution can be viewed as an infinite ensemble of point estimations $\boldsymbol{p}^{(i)}$. Therefore, EDL enables a better-calibrated way of quantifying *epistemic uncertainty* compared to traditional *Softmax*-based deep learning (Malinin and Gales, 2018; Sensoy et al., 2018). Additionally, the expectation of probability $\hat{\boldsymbol{p}}^{(i)}$ presents the average predictive confidence which reflects the *aleatoric uncertainty*. EDL is also able to capture the *distributional shift*. If no remarkable evidence can be modelled for a given input, the posterior $\alpha_c, \forall c \in C$ will approach 1, *i.e.*, the prior. Some illustrative examples of the posterior Dirichlet distributions are given in Figure 3.5.

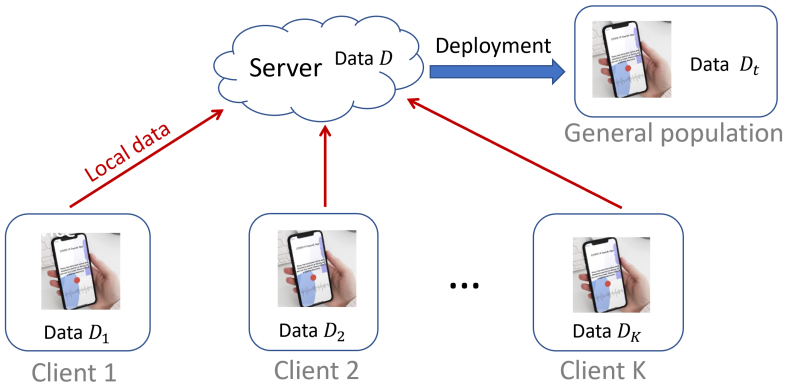
Relating to our work. DL-enabled health applications are designed for real-world deployment. They will have a positive impact on the health outcomes of the general population if they are properly used. Failures of those models may cause fatal consequences, therefore building trustworthy systems tends to be critical. Recently, studies have been attempting to incorporate uncertainty for human-in-the-loop medical diagnostics (Bhatt et al., 2021). For example, uncertainty estimation was applied to the task of diagnosing diabetic retinopathy from fundus images of the eye (Leibig et al., 2017; Van Amersfoort et al., 2020; Raghu et al., 2019). The quantified uncertainty can be used for selective prediction: keeping low-uncertain outputs but referring high-uncertain predictions to doctors, which includes clinicians in the loop and improves the system's robustness. Similarly, uncertainty-aware emotion recognition from video (Han et al., 2017), lung disease prediction from X-rays (Ghoshal and Tucker, 2020), and out-of-distribution detection for skin lesion diagnostic systems (Maron et al., 2021; Pacheco et al., 2020) are also studied. Beyond image models, Park *et al*. (Park et al., 2021; Xia et al., 2022a; Qendro et al., 2021a) benchmarked several uncertainty estimations and out-of-distribution detection methods on other data modalities including respiratory sounds, heart activity, brain waves, etc.

However, there is still a lack of comprehensive research dedicated to addressing the combined challenge posed by limited labelled data, class imbalance, and model overconfidence. In Chapter 4 and 5, we delve into ensemble learning and evidential deep learning, proposing innovative mechanisms to enhance their performance for reliable health diagnostics.

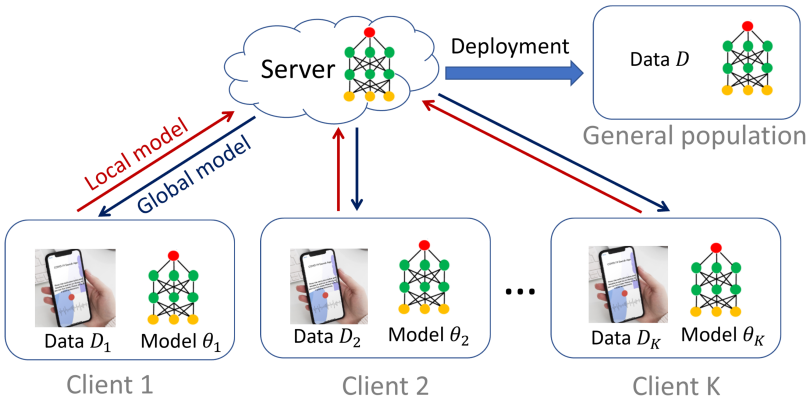
3.3.3 Federated learning for decentralised data

In the previous two sections, various techniques for enhancing the performance and reliability of deep neural networks were discussed. However, these training paradigms assume the existence of a centralised dataset $\mathcal{D}$, which is collected from multiple individuals or data collectors (see Figure 3.6(a)). Unfortunately, as identified as the third challenge in Chapter 1.2, privacy concerns arise when aggregating health data in a central place. This section provides an overview of a recently emerging technique known as federated learning (FL), which facilitates model training using decentralised data sources.

As illustrated in Figure 3.6(b), FL extracts knowledge from segregated data silos into a global model, avoiding the need to centralise the data in a single repository. Formally, starting with random initialisation, *i)* the global model is sent to clients, allowing them to optimise it using their local private data, denoted as $\mathcal{D}_k$, and then *ii)* the server gathers the updated parameters from clients and aggregates them into a new global model; Those two steps are repeated until the global model converges. The most representative and popular aggregation method is ***FedAvg*** (McMahan et al., 2017; Li et al., 2020b; Gao et al., 2022; Feng et al., 2022a), which averages the model parameters weighted by the fraction of local data sizes. Specifically, at round



(a) Model training using centralised data.



(b) Model training using distributed data.

Figure 3.6: **A comparison can be made between general model training using centralised data and federated learning using decentralised data**. In the centralised setting, data from multiple clients or users can be collected and used to train the model. In the decentralised setting, data remains locally with each client, and only model parameters are shared.

t, the global model $\theta^{(t)}$ is aggregated by,

$$\theta^{(t)} = \sum_{k \in \mathcal{K}^{(t)}} \frac{|\mathcal{D}_k|}{\sum_{k \in \mathcal{K}^{(t)}} |\mathcal{D}_k|} \theta_k^{(t)}, \tag{3.12}$$

where $\theta_k^{(t)}$ is the local model trained from private data $\mathcal{D}_k$, and $\mathcal{K}^{(t)}$ denotes the clients participated in this round.

This *FedAvg* method has demonstrated its ability to converge toward the optimal model, which is a model trained on the union of $\mathcal{D}$ and local datasets $\mathcal{D}_k$. This convergence is particularly evident when the local datasets $\mathcal{D}_k$ are sampled from the same distribution as $\mathcal{D}$, as shown in prior work (Li et al., 2019). However, in cases where $\mathcal{D}_k$ is drawn from different distributions, a scenario known as the ***data distribution heterogeneity*** (Zhao et al., 2018; Zhu et al., 2021; Li et al., 2020b; Luo et al., 2021), *FedAvg* tends to yield lower-performing models. Typically, $\mathcal{D}_k$ con-

tains different subsets of the globally available classes $\mathcal{C}$, each with varying numbers of samples, resulting in label distribution heterogeneity (Zhang et al., 2022; Luo et al., 2021). Notably, real-world physiological data distributed in different places often exhibit such heterogeneous distribution. For instance, in mobile health applications, individuals could exhibit different health conditions, rendering the local disease distribution unrepresentative of the global disease distribution at the population level. Effectively addressing the challenges posed by data heterogeneity is essential to enable the deployment of federated learning for privacy-aware health diagnostic models.

Relating to our work. While FL has shown significant promise in addressing privacy concerns when developing diagnostic models from personal data, it encounters substantial challenges, primarily attributed to the heterogeneous characterises of physiological data. We will delve into innovative mechanisms designed to tackle the challenges presented by class imbalance in Chapter 6 and label distribution skew in Chapter 7. Through these efforts, our study paves the way for the implementation of high-performing FL approaches in the healthcare domain.

3.4 Performance evaluation metrics

Throughout the studies in this thesis, we employ several evaluation metrics to demonstrate the performance of our approaches in terms of accuracy and uncertainty quantification. For diagnostics performance, we report key metrics, including ***AUC-ROC***, ***ACC***, ***Sensitivity***, and ***Specificity***. A larger value indicates better performance for these metrics. Regarding the quality of uncertainty estimates, we utilise metrics such as ***Brier*** and ***ECE***. Conversely, for both uncertainty metrics, a smaller value indicates better-calibrated uncertainty. By utilising this comprehensive set of evaluation metrics, we aim to provide a thorough analysis of our approaches, emphasising their accuracy in diagnostics tasks and the quality of uncertainty estimation.

In this thesis, for all clinical tasks, a data point is considered a positive case if a specific disease is present; otherwise, it is deemed negative. To facilitate the formulation of metrics for evaluating disease detection performance, we define the following terms to simplify the presentation,

- **TP** (True Positives): The number of cases having the disease, identified as having the disease.
- **FP** (False Positives): The number of cases not having the disease, identified as having the disease.
- **TN** (True Negatives): The number of cases not having the disease, identified as not having the disease.
- **FN** (False Negatives): The number of cases having the disease, identified as not having

the disease.

The true positive rate (TPR) is the number of cases having the disease and are identified as having the disease, divided by the total number of cases having the disease, *i.e.,* TPR = TP/(TP+FN). The false positive rate (FPR) is the number of cases not having the disease but are identified as having the disease, divided by the total number of cases which do not have the disease, *i.e.,* FPR = FP/(FP+TN).

Based on the above definitions, the metrics used for evaluations are introduced below.

ROC-AUC. As introduced in Chapter 3.1.2, the output of a classification ML model is a probability for a specific health condition. A threshold is needed to determine whether the prediction corresponds to having the disease or not. A receiver operating characteristic curve, commonly known as an ROC curve, is a graphical representation that depicts the classification performance of a binary classifier system as the discrimination threshold is adjusted. Simply speaking, the ROC curve illustrates the relationship between TPR and FPR at different thresholds. The TPR is the proportion of actual positive cases correctly identified as positive by the classifier, while FPR is the proportion of negative cases incorrectly classified as positive. The area under the ROC curve, denoted as ROC-AUC, quantifies the degree or measure of separability provided by the classifier, indicating how well the model can distinguish between different classes. A higher AUC value suggests that the model has a stronger ability to correctly predict instances of the positive class as positive and instances of the negative class as negative, signifying a more effective differentiation between positive and negative classes.

This metric is used to evaluate the overall performance for binary diagnostics, as in the studies presented in Chapter 4.3 and 6.3. It will also be employed to evaluate out-of-training-distribution detection in the study presented in Chapter 5.3.

ACC. ACC, short for *Accuracy*, presents the overall diagnostic performance at the categorical level. It measures the proportion of correct predictions made by a classification model out of all predictions made, and mathematically, it can be defined as ACC=(TN+TP)/(TN+TP+FN+FP). This metric is used for multi-class diagnostics experiments, as presented in Chapter 5.3 and 7.3.

Sensitivity. Sensitivity, also known as recall or TPR, is the proportion of actual positive cases correctly identified as positive by the classifier.

Specificity. Specificity, also known as the true negative rate (TNR), is the proportion of actual negative cases (the healthy control group) correctly identified as negative by the classifier, *i.e.,* TNR = TN/(TN+FP).

Sensitivity and *Specificity* are commonly used metrics in healthcare studies, and in our experiments, they consistently serve to showcase the performance of the proposed methods for accurate

health diagnostics. In binary diagnostics, such as those explored in Chapter 4.3 and 6.3, healthy controls are grouped into the negative class, while the unhealthy participants are considered the positive class. In the case of multi-class diagnostics, as examined in Chapter 7.3, we calculate these two metrics for each target class and then report their average. Unlike *ACC*, *Sensitivity* and *Specificity* are independent of the class distribution in the testing set, making them more reliable in most scenarios. While *ROC-AUC* considers all discrimination thresholds, when reporting *Sensitivity* and *Specificity* of a model, a single threshold is usually used. A commonly used threshold is 0.5, as the predictive probability ranges from 0 to 1. However, 0.5 may not be the best choice when the predicted probabilities are not calibrated and the distribution of classes is skewed. The threshold that aligns with the upper-left corner of the ROC curve, which maximises the difference between TPR and FPR (*i.e.*, balancing *Sensitivity* and *Specificity*), is often the optimal threshold for most cases. We will specify how the threshold is selected when we use these metrics in experiments.

Brier score. Following (Postels et al., 2022), the Brier score is used to measure the accuracy of predicted probabilities. Specifically, Brier score for a sample is computed as the squared error of a predicted probability vector, $\boldsymbol{p}^{(i)}$, and the one-hot encoded true response, $\tilde{\boldsymbol{y}}^{(i)}$, derived by $B^{(i)} = \frac{1}{C}\sum_{c=1}^{C}(\boldsymbol{p}^{(i)}[c] - \tilde{y}^{(i)}[c])^2$, for each sample. We report the average Brier score across the whole testing set, denoted by,

$$B = \frac{1}{C}\sum_{c=1}^{C}\frac{1}{N_c^t}\sum_{y^{(i)}=c} B^{(i)}, \tag{3.13}$$

where N_c^t denotes the number of samples from the c-th class in the testing set.

ECE. ECE, short for Expected Calibration Error, is defined to measure the correspondence between predicted probabilities and empirical accuracy (Ovadia et al., 2019). ECE quantifies the gap in the reliability diagram as shown in Figure 3.2. We used $M = 10$ universal bins to calculate ECE as follows:

$$\sum_{m}^{M}\frac{|B_m|}{N_{test}}|ACC(B_m) - conf(B_m)|, \tag{3.14}$$

where bin B_m covers the confidence interval $(\frac{m-1}{M}, \frac{m}{M}]$. $ACC(B_m)$ and $conf(B_m)$ are the ACC and the average predictive confidence for the samples having the predictive confidence within B_m.

Brier and *ECE* are used to validate the performance of model calibration and the quality of uncertainty estimates in the study present in Chapter 5.3.

Chapter 4

DB-EL: Uncertainty-aware deep learning for binary physiological data

Knowledge is an unending adventure at the edge of uncertainty.

- Jacob Bronowski

Polish-British mathematician and philosopher

4.1 Introduction

Employing deep learning to model health shows great promise in the literature; however, concerns remain regarding its reliability in real-world health diagnostic settings (Park et al., 2021). These concerns primarily arise from the observations that deep learning models can often exhibit poor calibration, leading to overconfident probabilities that do not accurately reflect true confidence levels (Guo et al., 2017). As introduced in Chapter 1.2, overconfident yet incorrect model predictions can result in unacceptable costs in healthcare. Therefore, calibrating deep learning models for health diagnostics is an essential task.

As discussed in Chapter 3.3.2, ensemble learning serves as an approach to quantify model uncertainty by employing a finite number of models (Ganaie et al., 2022; Dietterich, 2000; Zhou, 2012). Ensemble deep learning, which integrates multiple deep neural networks, usually leads to a better-calibrated model when compared to the traditional single *Softmax*-based neural network. In this chapter, we explore ensemble deep learning for health screening based on physiological data, which is often characterised by small-scale and severe class imbalance (as introduced in Chapter 1.2).

Such data characteristics not only exacerbate the model's overconfidence but also introduce bias into the model. Since the model may prioritise optimisation for the majority class (*i.e.*, the healthy control group), both the classification boundary and the confidence can be biased, becoming inaccurate for the minority class (*i.e.*, the unhealthy group). To address this challenge, this chapter proposes a novel data-balanced deep ensemble learning approach (***DB-EL***) for reliable binary health screening. In this approach, we generate multiple balanced raining sets from imbalanced physiological data through a re-sampling strategy. Those sets are then leveraged to train multiple model ensemble units. Predictions from these units are fused to calibrate the confidence from a single model. In addition, the inconsistency among the predictions of the learned units is used as a measure of model uncertainty. This uncertainty quantification can enhance the reliability of the model when deployed to the real world.

This chapter makes the following contributions,

- We introduce a novel ensemble learning method designed to construct a well-calibrated binary health screening model using limited and class-imbalanced physiological data.
- We conduct a study utilising physiological audio to predict respiratory health conditions, specifically, whether an individual is COVID-19 positive or not. The results demonstrate the superior accuracy of our method compared to baselines, showing an improvement of 7.2% in ROC-AUC. Additionally, our method exhibits better calibration than a single deep-learning model.
- In this respiratory health screening task, we show that the quantified uncertainty can effectively indicate the correctness of model predictions. Therefore, by leveraging uncertainty measurements for selective prediction, we achieve an additional screening accuracy boost of 17.6%.

The remainder of this chapter is organised as follows. We first review the related studies in Chapter 4.2. Then, we introduce our method in Chapter 4.3. The experimental setup and results are presented in Chapter 4.4 and 4.5, respectively. We finally conclude our findings in Chapter 4.6.

4.2 Related work

As discussed in Chapter 3.3.2, conventional *Softmax*-based deep classifiers are limited in uncertainty quantification as they only generate deterministic point estimations. Ensemble learning, a frequentist method for uncertainty estimation, involves training multiple models using different subsets of the data, model initialisation, or model architectures (Ganaie et al., 2022). This chapter explores deep ensemble learning in audio-based respiratory health screening. The most

related studies are reviewed as follows.

Ensemble learning for health applications. Ensemble learning has found widespread application in health diagnostics to enhance model accuracy (Sarmadi et al., 2021; Raza, 2019; Nguyen et al., 2021b; Cao et al., 2020). For example, Raza *et al.* proposed the utilisation of ensemble learning to enhance classification accuracy in heart disease detection (Raza, 2019). This study incorporated risk factors such as age, gender, heart rate variability, blood pressure, cholesterol, and blood sugar, among others, into multiple machine learning models, including decision trees, support vector machines, and k-nearest neighbours. The integration of different models can improve performance by capturing heterogeneous patterns from the data. In the realm of deep learning for health diagnostics, ensemble learning is also commonly used via probability-wise fusion to calibrate diagnostic confidence and quantify model uncertainty (Nguyen et al., 2021b; Leibig et al., 2017; Raghu et al., 2019; Zheng et al., 2023). In respiratory health, due to the limited availability of audio data for model training, ensemble learning is also widely used to boost model performance. Nguyen *et al.* proposed a snapshot ensemble learning method, which combines deep neural networks trained in several epochs (Nguyen and Pernkopf, 2020). This method reduces the cost of training different models and mitigates the error of a single model snapshot.

Combination methods. Since ensemble learning involves a number of models, determining how to combine those base models to produce the final prediction has been a long-standing problem. Unweighted averaging of the outputs of the base models in an ensemble is the most commonly followed approach for fusing decisions in the literature (Ganaie et al., 2022). Here, the outcomes of the base models are averaged to obtain the final prediction of the ensemble model. Deep learning architectures exhibit high variance and low bias; thus, simple averaging of the ensemble models improves generalisation performance by reducing variance among the models. Unweighted averaging is a reasonable choice when the performance of the base models is comparable, as suggested in (He et al., 2016). However, it has also been recognised that when the ensemble contains heterogeneous base models, naive unweighted averaging may result in sub-optimal performance, as it is affected by the performance of the weak base models and the overconfident base models (Ju et al., 2018).

To overcome the limitation of unweighted averaging, other voting methods have been investigated. Similar to unweighted averaging, majority voting combines the outputs of the base models. However, instead of taking the average of the probability outcomes, majority voting counts the votes of the base models and predicts the final labels as the label with the majority of votes. The majority voting technique was employed to improve diagnostic accuracy for the ensemble method compared to individual classifiers (Raza, 2019). The predictions of base models can also be integrated by the Bayes optimal classifier, where the prediction of each base model is regarded as the conditional distribution of the target label. Choosing prior probabili-

ties in the Bayes optimal classifier is difficult and hence is usually set to a uniform distribution for simplicity. With a large sample size, one hypothesis tends to give larger posterior probabilities than others, and hence the weight vector is dominated by a single base model, causing the Bayes optimal classifier to behave as the discrete super learner with a negative likelihood loss function (Ganaie et al., 2022). More recently, Bayesian non-parametric methods have been studied in ensemble learning for better voting: a deep neural network is used to map inputs into a latent feature space, where a Gaussian process with a base kernel acts; the resulting model is then trained in an end-to-end fashion (Liu et al., 2018; Ober et al., 2021). This approach can outperform the common ensemble where base models are randomly initialised and trained independently, as the base models tend to be more diverse (*i.e.*, less correlated with one another).

Uncertainty-aware diagnostics. The uncertainty quantification capability inherent in ensemble learning not only enhances the accuracy of disease detection but also fosters the integration of a collaborative approach between humans and machines. By leveraging this capability, medical professionals gain valuable insights into cases where deep learning models encounter uncertainty (*i.e.*, models do not know the case), indicating potential instances of misdiagnosis. This collaboration between human expertise and machine intelligence enables timely intervention and ensures that critical cases receive the attention they require. Additionally, the ability to quantify uncertainty empowers healthcare practitioners to make informed decisions, particularly in complex or ambiguous situations where the model's confidence may be compromised. As a result, the collaborative synergy between human expertise and ensemble learning models not only enhances diagnostic accuracy but also augments the overall efficacy and reliability of healthcare systems (Bhatt et al., 2021; Pacheco et al., 2020; Kang et al., 2021). For example, deep ensemble learning has been applied to diagnose diabetic retinopathy from fundus images of the eye (Leibig et al., 2017; Raghu et al., 2019). Quantified uncertainty can guide selective predictions by retaining low-uncertainty outputs and referring high-uncertainty predictions to clinicians, thereby involving clinicians in the decision-making process and enhancing the system's robustness.

Ensemble Learning for Class Imbalance. Although ensemble learning has been extensively studied, as pointed out in a review (Cao et al., 2020), leveraging ensemble learning to address the common challenge of class imbalance is largely under-explored. In contrast to existing studies, our work uniquely focuses on the prevalent issue of data imbalance, a concern often overlooked in uncertainty quantification research. Van *et al.*'s study identified the adverse impact of class imbalance, specifically noting that the quality of uncertainty is compromised, particularly when distinguishing incorrect predictions for the minority class (down-sampled to 5$\sim$10% of the original data size) (Van Molle et al., 2021). In light of this, our method is designed to mitigate the model bias and overconfidence stemming from the constraints of limited and imbalanced physiological data. We address this challenge within a unified framework, seeking to enhance the

robustness and reliability of uncertainty quantification in the face of class imbalance.

4.3 Methodology

4.3.1 Problem formulation

In this section, we focus on model calibration and uncertainty quantification for binary health screening applications and propose a novel solution to address the imbalance issue in physiological data. For the sake of clarity, we first formulate the problem below. Following that, we introduce our proposed method.

***Uncertainty-aware deep learning for binary health screening**: Consider a physiological dataset with two classes denoted as $\mathcal{D} = \{(x^{(i)}, y^{(i)})\}_{i=1}^{N}$, where $x^{(i)}$ is a physiological data sample, and $y^{(i)}$ represents the health condition. That is, if the associated disease is identified in the sample, $y^{(i)} = 1$ (unhealthy); otherwise, $y^{(i)} = 2$ (healthy). Here, N is the number of training samples, and y is extremely imbalanced, with $y = 1$ being the minority class. The task is to train a deep learning model parameterised by θ that can predict y for any given x to achieve population health screening. We aim for calibrated predictions, and for each prediction, an uncertainty measurement is provided.*

4.3.2 Data-balanced ensemble learning

> ***Highlight**: Ensemble learning can integrate predictions from multiple models to mitigate the issue of overconfident predictions made by a single model. Training each model within the ensemble using a balanced subset of the entire training set is supposed to not only enhance the utility of the data but also reduce model bias, thereby leading to reliable uncertainty quantification for imbalanced data.*

To tackle the above-defined problem, we propose a data re-sampling strategy to optimise the utilisation of such physiological data. We hypothesise that ensemble learning can alleviate the overconfidence inherent in individual models. Concurrently, our data re-sampling strategy is designed to tackle the challenge of class imbalance. An overview of our method is depicted in Figure 4.1, with each component introduced as follows.

Model training. Firstly, we re-sample the heavily imbalanced training set to create N_m balanced subsets, denoted as $(X_1, Y_1), (X_2, Y_2), ..., (X_{N_m}, Y_{N_m})$. Each subset (X_n, Y_n) comprises the data from an equal number of healthy and unhealthy participants, where X_n is a collection of physiological samples $x^{(i)}$, and Y_n is the collection of the corresponding health conditions $y^{(i)}$. Notably, the unhealthy group typically constitutes the minority class. To ensure the com-

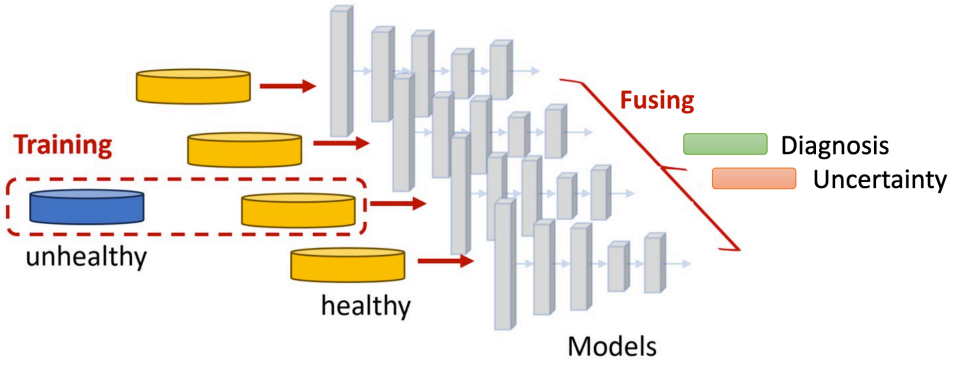


Figure 4.1: **Data-balanced ensemble deep learning for health screening**. Balanced training sets are generated to train multiple models, and probabilities for a testing sample are fused to form the final decision. Simultaneously, the disagreement level across these models as a measure of uncertainty is obtained and used to indicate the reliability of digital diagnoses.

prehensive utilisation of these samples, we incorporate them into each subset (indicated by the blue box in Figure 4.1), while randomly selecting healthy samples (as depicted by the yellow box in Figure 4.1). Consequently, we independently train N_m models using the N_m subsets.

These balanced subsets are employed for training deep learning models independently for health screening. In this study, we adapt the same architecture for each model. The training process follows the back-propagation optimisation method introduced in Chapter 3.1.2. The models within this ensemble framework may have similar or different architectures, depending on the specific health condition tasks.

Model fusing. Once all the models are trained, any testing sample $x^{(i)}$ can be fed into the ensemble for predictions. For the final output, we use the probability-wise prediction averaging, formulated as follows,

$$\boldsymbol{p}^{(i)} = \frac{1}{N_m} \sum_{n=1}^{N_m} \boldsymbol{p}_n^{(i)}, \tag{4.1}$$

where $\boldsymbol{p}_n^{(i)}$ is the predicted *Softmax* probability from the n-th model. The expectation $\boldsymbol{p}^{(i)}$ is more calibrated than any $\boldsymbol{p}_n^{(i)}$ as the occasional model overconfidence can be smoothed. For the final prediction, if $\boldsymbol{p}^{(i)}[1]$ is greater than $\boldsymbol{p}^{(i)}[2]$, the testing sample i will be predicted as belonging to the unhealthy class; otherwise, it will be predicted as healthy.

This model fusion technique known as unweighted averaging is commonly employed. Additionally, various other methods for model fusion, as discussed in Chapter 4.2, can be seamlessly integrated into our framework.

Uncertainty estimation. In a clinical context, when multiple clinicians provide a diagnosis for a patient if their conclusions are consistent, it indicates a low level of uncertainty in the

diagnosis. Conversely, when there is disagreement among clinicians, it suggests that the case is more challenging, and the confidence in the diagnosis decreases (Schumacher et al., 2013; Raghu et al., 2019). Similarly, we use the inconsistency among predictions from multiple models as an indicator of the model's uncertainty. Formally, we use the standard deviation σ of the predicted likelihood for the positive class across the N_m models as the measurement of uncertainty as follows,

$$\begin{aligned} \mu^{(i)} &= \frac{1}{N_m}\sum_{n=1}^{N_m}(\boldsymbol{p}_n^{(i)}[1]), \\ \sigma^{(i)} &= \sqrt{\frac{1}{N_m}\sum_{n=1}^{N_m}(\boldsymbol{p}_n^{(i)}[1]-\mu^{(i)})^2}. \end{aligned} \tag{4.2}$$

If the uncertainty $\sigma^{(i)}$ is higher than a predefined threshold, it implies that the model is unsure of its prediction during digital screening. Under this circumstance, the system can first request a second or even more repeated audio testing on smartphones. If the uncertainty is still high, this particular sample could be then referred for further clinical or other testing. As a consequence, both system performance and patient safety can be improved.

Overall, the probability-wise fusion (Eq. (4.1)), based on the learned multiple models, can mitigate the overconfident predictions of a single model. Furthermore, the variance (Eq. (4.2)) among the predictions made by the model ensembles provides explainable measures of model uncertainty.

4.4 Experimental setup

4.4.1 Dataset

To evaluate the proposed framework, we utilise the *COVID-19 Sounds* database (as introduced in Chapter 2.2.1) for experiments. In this study, we include participants who tested positive and exhibited at least one symptom, as well as participants who tested negative and declared no symptoms. To eliminate language confounders in the voice recordings, only English speakers are retained. Ultimately, we include 330 positive participants with 469 samples and 919 negative participants with 2,021 samples. Consequently, the dataset is small and heavily imbalanced, which is suitable for our evaluation. Overall, 58% of the participants are male, and more than 60% are aged between 20 and 49. Demographics and medical history distributions are similar in the two classes.

The task involves predicting the COVID-19 status of a given sample. To accomplish this, we develop a deep learning model, as detailed in Chapter 4.4.2. For training and evaluation, we divide the data into three sets. Specifically, for the positive group, we set aside 10% and 20% of partic-

Table 4.1: **Basic statistics of *COVID-19 Sounds* data used in this study**. The data presents class imbalance at both participant and sample levels.

	Positive		**Negative**	
	#Participants	*#Samples*	*#Participants*	*#Samples*
Training set	231	327	820	1,871
Validation set	33	44	33	56
Testing set	66	98	66	94

ipants for validation and testing, respectively, using the remaining data for training. This results in 231 positive participants for model training. Correspondingly, we select the same number of negative participants for validation and testing, leading to 820 negative participants for training. The statistics of the three sets are summarised in Table 4.1. To generate balanced subsets for training model ensembles, 231 participants are randomly selected from the 820 negative tested participants for each subset.

Regarding data pre-processing, we re-sample all the recordings to 16 kHz mono audios, removing the silence period at the beginning and end of each recording. Finally, audio normalisation, achieved by calibrating the peak amplitude to 1, is applied to eliminate discrepancies across recording devices. We set $N = 10$ so that 10 models are learned. During training, our batch size is 1, the learning rate is 0.0001 with a decay factor of 0.99, and we use cross-entropy loss and the Adam optimiser. Early stopping is applied to the validation set to obtain the best model for reporting performance.

4.4.2 Backbone model and training strategy

As introduced in Chapter 2.2.1, the spectrogram is the commonly used representation for audio waves since it can effectively capture both temporal and frequency features. Herein, we apply the STFT to derive spectrograms for the respiratory audio samples in our experiments. Since a spectrogram is a two-dimensional input, we adopt a CNN-based model architecture for this COVID-19 screening task (Ren et al., 2020). The overall framework, as illustrated in Figure 4.2, comprises the following components:

Spectrogram input. As introduced in Chapter 2.2.1, audio samples, before being fed into deep learning models, are usually transformed into spectrograms via STFT. In this task, each single audio recording is initially divided into non-overlapping segments of 960 ms each. STFT is then applied to each segment using a window length of 25 ms with 10 ms overlap, and a periodic Hann window is used. This process results in the creation of a spectrogram. Furthermore, the spectrogram is integrated into 64 Mel-spaced frequency bins, and the magnitude of each bin is log-transformed after adding a small offset to avoid numerical issues (Ali et al., 2021). As a

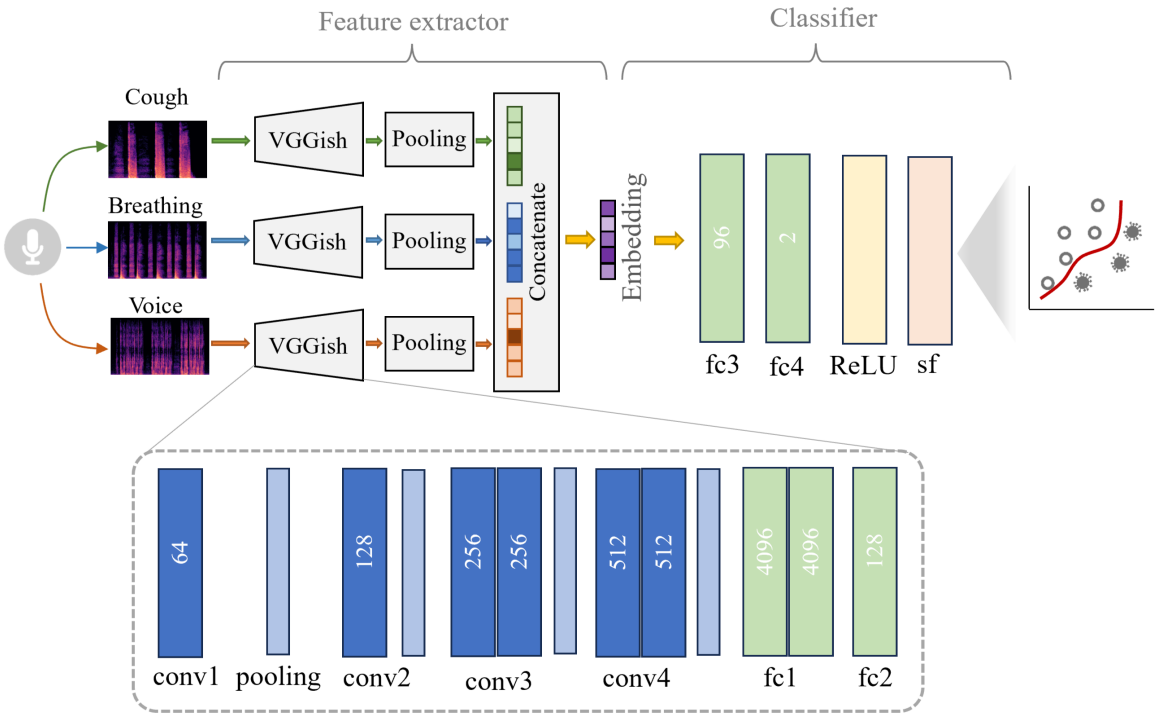


Figure 4.2: **Architecture of our CNN model.** The model consists of the feature extractor and the classifier parts. VGGish is employed to extract acoustic features from cough, breathing, and voice recordings. These features are then concatenated into a single embedding vector. These features are subsequently concatenated into a single embedding vector for classification. In this VGGish model, the blue, light blue, and green blocks represent convolutional, pooling, and fully connected layers, respectively.

result, the final spectrogram for each segment has dimensions of 96×64.

Feature extractor. Each spectrogram is subsequently processed with the VGGish module to extract deep features (Simonyan and Zisserman, 2014; Hershey et al., 2017). VGGish is a convolutional neural network-based architecture comprising cascaded convolutional layers, max-pooling, and fully connected layers. For each 960 ms audio segment, VGGish can transform its spectrogram into a 128-dimensional embedding vector (known as features in deep learning). The architecture of VGGish is illustrated in Figure 4.2, and the components are summarised below:

- Convolutional Layers: VGGish contains several convolutional layers with 3×3 filters. These layers learn to extract hierarchical features from the input Mel spectrogram patches. As shown in Figure 4.2, the number of filters in each convolutional layer gradually increases as going deeper into the network.

- Max Pooling Layers: After each convolutional layer, VGGish applies max pooling with a 2×2 window and stride of 2. This reduces the spatial dimensions of the feature maps

while preserving important features.

- Fully Connected Layers: Following the convolutional and max pooling layers, VGGish has two fully connected layers with *ReLU* activation functions (as introduced in Chapter 3.1.2). These layers perform non-linear transformations on the extracted features to capture higher-level representations.

- Embedding Layer: The final layer of VGGish is an embedding layer that produces a 128-dimensional embedding vector for each input Mel-spectrogram patch. This embedding vector represents the learned features of the input audio segment and serves as the output of the network.

The embedding vectors of segments from the same audio sample are aggregated to form the feature for that sample. Following the VGGish model, an average pooling layer is employed to combine the embedding vectors derived from all segments within a particular audio recording. This results in the creation of a fixed-length latent feature vector, regardless of the individual recording's duration. Finally, the resulting embedding vectors for the three modalities (cough, breathing, and voice) are concatenated to form a multi-modal embedding vector. This embedding vector is then used to classify each audio sample.

Considering the small scale of the available training data, we explore transfer learning techniques as introduced in Chapter 3.1.2. We leverage the VGGish model, which was pre-trained on a public benchmark dataset. This benchmark dataset comprises 100 million YouTube audio recordings totalling 5.4 million hours (Hershey et al., 2017). Therefore, the pre-trained VGGish exhibits good acoustic feature extraction capabilities. Since the VGGish was not specifically pre-trained for respiratory sounds, we also update its parameters when training the other components of the model, as introduced below.

Classifier. The extracted embedding vector from the sample is inputted into a binary classifier as depicted in Figure 4.2. This classifier comprises two fully connected layers (abbreviated as *fc*) with non-linear *ReLU* and a *Softmax* layer (as introduced in Chapter 3.1.2). The number of hidden states in these fully connected layers is 96 and 2, respectively. Before the two *fc* layers, Dropout with a probability of 0.5 is applied to avoid overfitting. The model's output is a two-dimensional probability vector $\boldsymbol{p}^{(i)}$, indicating the probabilities of being positive or negative. Unless otherwise specified, we consider the categorical prediction as the class with the higher probability.

To fit the model parameters, including the parameters for VGGish and the binary classifier, the following binary cross-entropy loss for each sampling in the training set is utilised,

$$\mathcal{L}^{(i)} = \log \boldsymbol{p}^{(i)}[y^{(i)}], \tag{4.3}$$

where the label $y^{(i)} = 1$ when sample $x^{(i)}$ is a COVID-19 positive case, otherwise $y^{(i)} = 2$. As previously explained in Chapter 3.1.2, the optimisation of model parameters, which encompasses both the feature extractor and the classifier, can be achieved through the utilisation of Stochastic Gradient Descent (SGD). This involves the back-propagation of the averaged loss over a small batch of the training samples.

4.4.3 Baselines and metrics

In addition to deep models, acoustic feature-driven classifiers are reported to achieve state-of-the-art performance in sound-based COVID-19 detection, due to their effectiveness and robustness in small data learning (Brown et al., 2020a; Han et al., 2020, 2021a). Therefore, we use the method in (Han et al., 2020), named **SVM**, as our baseline. This method leverages the openSMILE toolkit to extract acoustic features (Eyben et al., 2010), and SVM as the classifier. OpenSMILE (*open-source Speech and Music Interpretation by Large Space Extraction*) is a widely used software package for extracting acoustic features from audio signals. It is commonly employed in various applications such as speech processing, emotion recognition, and speaker identification. It provides a comprehensive set of low-level and high-level acoustic features that capture different aspects of the audio signal. These features include:

1. Low-level features: These features are derived directly from the audio waveform and include parameters such as pitch, loudness, and spectral shape.

2. High-level features: These features are derived from the low-level features and represent more abstract characteristics of the audio signal, such as speech prosody, voice quality, and emotion content.

As a result, a total of 384 acoustic features are fed into the SVM model. PCA (*Principal Component Analysis*) is used to reduce the dimensionality of the features. We finally retain the features that explain 90% of the covariance in the data.

For both SVM and deep models, we compare training **a single model** with training $N_m = 10$ models for the ensemble. Using all samples and also exploring balanced datasets created through down-sampling or up-sampling. In the case of down-sampling, we randomly discard some negative samples, while for up-sampling, we employ the Synthetic Minority Over-sampling Technique (SMOTE) (Chawla et al., 2002) to generate synthetic recordings for the positive class.

To demonstrate that our ensemble learning method can yield good model confidence estimation, we compare DB-EL with *MCDropout*, the uncertainty quantification method as detailed in Chapter 3.3.2. Since our CNN model is trained with dropout layers, to test this baseline, we keep the dropout enabled and pass each sample into the model 10 times to obtain the variance as defined in Eq.(4.2). We are unable to compare our DE-EL approach with Bayesian neural net-

works as introduced in Chapter 3.3.2, since the VGGish has been pre-trained as a deterministic model.

To validate the performance of the proposed framework for COVID-19 screening, we report the following metrics: ***ROC-AUC***, ***Sensitivity***, and ***Specificity*** (as detailed in Chapter 3.4). Furthermore, for both the baseline and our proposed methods, we report the mean and standard deviation of these metrics across 10 runs by using different random seeds.

The model parameters are optimised using the training set, and the hyper-parameters are determined through the validation set, as summarised in Table 4.1. For the SVM, we explore different kernels (options include 'linear', 'poly', 'rbf', and 'sigmoid') and the regularisation parameter C (choosing from 0.01, 0.1, 1.0, 10, 100)[1]. For our CNN model, we investigate the size of the classifier (selecting the number of neurons from 32, 64, 128, 256) and the learning rate (options are 0.0001, 0.001, 0.01, 0.1). We ultimately select the hyper-parameters that yield the best ROC-AUC score on the validation set. As a result of this search, for the SVM model, we use a linear kernel and set the regularisation constant C to 0.01. For the CNN model, we use 96 neurons for the *fc3* layer and a learning rate of 0.001. The operating point on the ROC curve, to report sensitivity and specificity, involves choosing a specific threshold value that determines how the model's predictions are classified into positive and negative outcomes. This threshold for each method is also identified by the validation set: we choose the threshold that minimises the distance to the top-left corner of the ROC plot on the validation set (Attia et al., 2019).

4.5 Results

4.5.1 Classification performance

The results for COVID-19 screening based on the insufficient and imbalanced training data are presented in Table 4.2. From the results, we have the following observations:

- **Deep learning is not superior to traditional machine learning with imbalanced training data**. From the first row of Table 4.2 (SM imbalanced data), we observe that although the CNN model achieves a ROC-AUC of 0.69, surpassing the SVM's 0.602, the sensitivity and specificity do not exhibit significant improvement.
- **Re-sampling can enhance performance**, particularly in sensitivity, for both SVM and deep learning models, as it ensures a balanced training set. A comparison of the second and third rows of the table with the first row reveals that, in contrast to down-sampling, up-sampling yields superior performance compared to both down-sampling and no-sampling.

[1] We adapt the implementation from https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html#sklearn.svm.SVC

Table 4.2: **Performance comparison.** We report Mean±Std for ROC-AUC, Sensitivity, and Specificity reported for the Single model (SM) and the Ensemble model. Optimal threshold is used to balance the sensitivity and specificity.

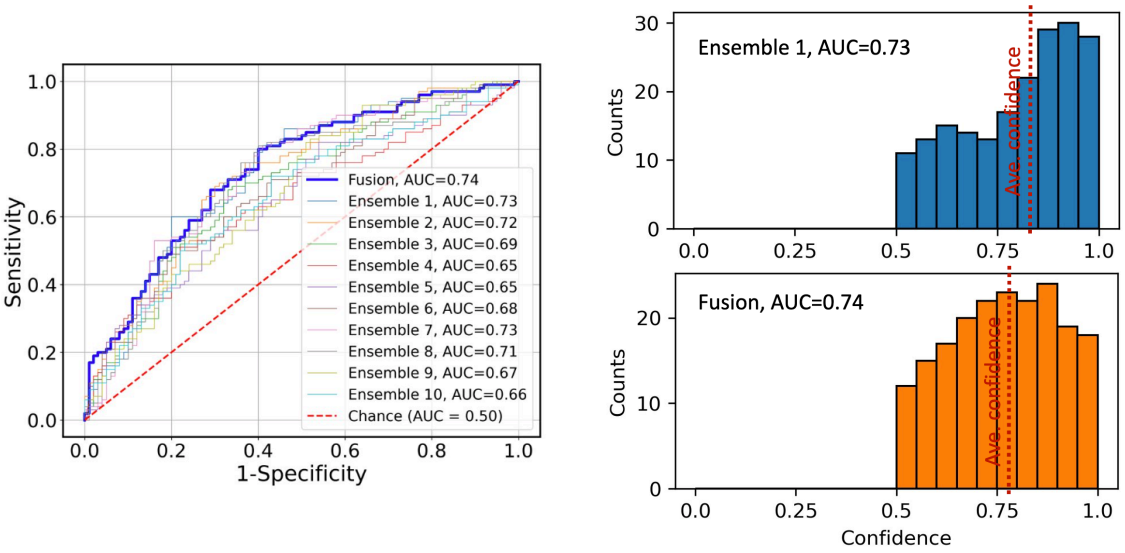
		ROC-AUC	Sensitivity	Specificity
SM imbalanced data	**SVM**	0.602	0.562	0.563
	CNN	0.690	0.571	0.572
SM down-sampling	**SVM**	0.601 ± 0.030	0.565 ± 0.052	0.565 ± 0.050
	CNN	0.682 ± 0.041	0.638 ± 0.043	0.643 ± 0.062
SM up-sampling	**SVM**	0.622 ± 0.021	0.585 ± 0.022	0.592 ± 0.021
	CNN	0.701 ± 0.039	0.675 ± 0.025	0.677 ± 0.051
Ensemble model re-sampling	**SVM**	0.660 ± 0.042	0.632 ± 0.053	0.624 ± 0.042
	CNN	0.740 ± 0.029	0.686 ± 0.051	0.689 ± 0.059

This is reasonable due to the increased availability of data samples for parameter learning.

- **Ensembles can enhance the performance of both SVM and deep model**. As shown in the last row of the table, ensemble models consistently improve the ROC-AUC compared to a single model, regardless of the data re-balancing strategy.
- **Our ensemble method outperforms all the baselines**. Our CNN model with ensembles yields a ROC-AUC of 0.740 with a sensitivity and a specificity close to 0.7. Compared to the highest ROC-AUC of 0.70 from the baselines, our method achieves a relative improvement of 5.7%. This demonstrates the superior accuracy of deep ensemble learning for COVID-19 screening from imbalanced data.

To further demonstrate the superiority of ensemble deep learning, we visualise the ROC curves for each model in Figure 6.3 for comparison. All ROC curves are above the chance level, but the model variance is not negligible. A plausible explanation is that we only have a small training dataset for each model, and it is reasonable to quantify the model uncertainty from the variance of these units. Additionally, it can be observed that after probability-based fusion, the ROC curve is generally higher than the other curves, yielding the highest ROC-AUC of 0.740.

We observe that the fusion not only enhances the accuracy of COVID-19 screening but also mitigates the model's overconfidence. Figure 6.3 illustrates the confidence distribution for the best single model unit and the model fusion. For a well-calibrated deep learning model, its predictive confidence typically aligns with its classification accuracy Guo et al. (2017). In Figure 6.3, the best single model (Ensemble 1) achieves a comparable ROC-AUC score to the fusion, but its average confidence of 0.83 is significantly higher than that of the fusion (0.79). Our model fusion generates a more uniform distribution across various confidence levels (for binary classification, confidence ranges from 0.5 to 1), while the confidence for the Ensemble is skewed toward the high-confidence range (over 0.75). From these findings, we can conclude that our data-balancing ensemble learning approach achieves a more calibrated COVID-19 screening



(a) ROC curves.

(b) Confidence histogram.

Figure 4.3: **Comparing the ensemble learning model to individual models**. (a) shows the ROC curves for each ensemble and the fusion, respectively. (b) shows the confidence distribution for the best ensemble unit and the fused model.

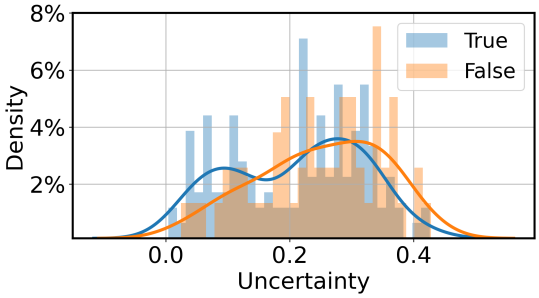
model compared to an individual model.

4.5.2 Uncertainty quantification performance

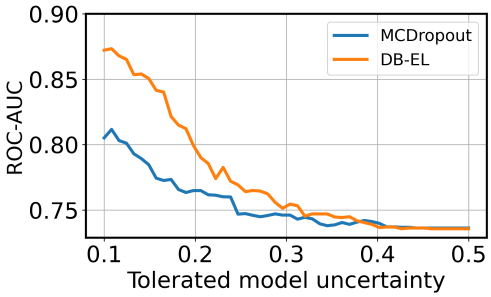
Now, let us examine the quality of our model uncertainty measurements and provide insights into how the ensemble learning approach can bolster the reliability of health diagnostics through uncertainty quantification.

To gain an initial understanding of the quality of our uncertainty measurements, we present the uncertainty distribution in Figure 4.4(a). Notably, the density of high uncertainty values for incorrect predictions (indicated by the False group) exceeds that of correct predictions (True group). This implies that our approach excels in discerning less confident predictions, particularly when an erroneous diagnosis is made.

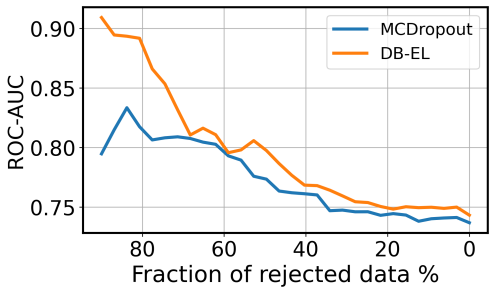
Motivated by these findings, we examine the performance of selective prediction by establishing thresholds to exclude certain testing cases where the uncertainty exceeds a specified value. The outcomes of our DB-EL method compared to the MCDropout baseline are depicted in Figure 4.4(b). By retaining only testing samples with uncertainty below 0.2, the ROC-AUC of our method improves significantly from 0.74 to 0.79, marking a 6.8% enhancement. Furthermore, at a threshold of 0.1, the highest ROC-AUC value of 0.87 is achieved, reflecting a substantial 17.6% improvement. Although the performance of MCDropout also increases with selective prediction (from 0.74 to 0.81), our method demonstrates a more notable boost in performance



(a) Uncertainty distribution.



(b) ROU-AUC.



(c) ROU-AUC.

Figure 4.4: **Performance for uncertainty quantification**. (a) shows the distribution of uncertainty for correct and incorrect prediction, respectively. (b) and (c) present the ROC-AUC for selective prediction with different uncertainty thresholds and percentage of rejection.

as shown in Figure 4.4(b). This indicates that our uncertainty estimates are more informative and useful.

These results suggest the significance of studying predictive uncertainty and identifying optimal thresholds. While referral for further clinical testing can enhance diagnostic accuracy, it also imposes an additional burden on doctors. To strike a balance, we examined the ROC-AUC across different fractions of rejected data (excluded from model predictions) with uncertainty above specific thresholds (assuming samples with uncertainty above these thresholds should be referred to doctors). Figure 4.4(c) shows that by excluding the 40% of samples with the highest uncertainty (threshold set at 0.28), the ROC-AUC increases from 0.74 to 0.77. These findings indicate that selectively directing an acceptable proportion of screening cases to doctors can achieve a balance between effectiveness and efficiency. Moreover, limiting the analysis to the remaining 20% of data with the lowest uncertainty (threshold set at 0.08) results in an increase in ROC-AUC to 0.89. Despite the great performance, we acknowledge that passing 80% of the samples to doctors would yield a heavy workload and is not practical.

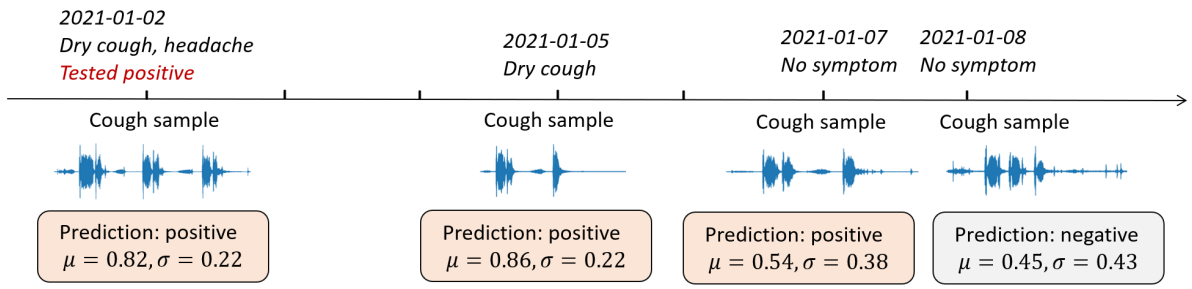


Figure 4.5: **Case study for uncertainty estimation.** This participant exhibited a progression toward recovery. The cough audio sample used and the model predictions are shown below the time axis.

4.5.3 Case study

To further demonstrate that introducing uncertainty quantification provides additional information for clinical decision-making beyond categorical model predictions, we present a case study of a participant in our testing set who provided multiple audio samples on different days. This participant is a male in his 20s, an English speaker, with no significant medical history. He tested positive for COVID-19 on January 2nd, presenting with severe symptoms. By January 5th, his headache had dissipated, and by January 7th, all symptoms had resolved. This participant contributed four samples, which were analysed using our model to predict his COVID-19 status. The results are presented in Figure 4.5.

For the first two audio samples, the model yielded a positive prediction with moderate confidence (the uncertainty measurement $\sigma = 0.22$). For the third sample, corresponding to the period when his headache had disappeared, the model continued to predict a positive outcome, but the uncertainty increased from 0.22 to 0.38. For the fourth sample, the model's prediction shifted to negative, yet the uncertainty level was high at 0.43, indicating the model's hesitance about the prediction. Considering the typical infectious period of 7 to 10 days during the first wave of the pandemic in the UK in 2021, this participant's recovery trend (possibly due to medical intervention) suggests it is reasonable for the model to become uncertain in its predictions based on audio data alone. If the COVID-19 prediction outcome is used to determine whether this participant can end self-quarantine, the level of uncertainty suggests the necessity for a confirmatory medical test.

In summary, these results demonstrate that our model's uncertainty measurement is an informative indicator of the correctness of model predictions, thereby playing an instrumental role in enhancing the reliability of an automated health screening system.

4.6 Discussion and conclusions

In this chapter, we showcased the effectiveness of deep ensemble learning in calibrating model predictions and providing high-quality uncertainty estimates, achieved through a data-balancing strategy. Notably, our study represented a novel exploration into uncertainty quantification on imbalanced physiological data. When applied to real-world physiological audio data, characterised by substantial class imbalance, our experiments demonstrated the superiority of our ensemble method over a single model without uncertainty quantification.

However, it is crucial to acknowledge the limitations associated with this ensemble approach. One limitation of our study lies in its focus on binary healthy diagnostics. While our data-balancing strategy could be extended to multi-class physiological data, we did not systematically evaluate this aspect, leaving it as a subject for future exploration. This ensemble learning approach also necessitates significant training resources and results in inference inefficiency, since multiple models need to be stored. This potentially poses a bottleneck for mobile health applications, given that many mobile devices may lack the capacity to accommodate multiple models.

Chapter 5

CB-EDL: Uncertainty-aware deep learning for multi-class physiological data

Information is the resolution of uncertainty.

- Claude Shannon

Father of information theory

5.1 Introduction

Accurate and efficient uncertainty estimation is of paramount importance in safety-critical applications such as mobile health diagnostics, where reliable uncertainty estimates are essential and computational resources are constrained (Qendro et al., 2021a; Rock et al., 2021; Qendro et al., 2021b; Steinhubl et al., 2015). As illustrated in Figure 3.4 in Chapter 3.3.2, prevalent uncertainty-aware deep learning methods, including Bayesian neural networks, ensemble deep learning, and MCDropout, either require the storage of additional model parameters or demand multiple forward passes for uncertainty estimation, falling short of meeting the efficiency requirements. In contrast, the evidential deep learning (EDL) technique, as detailed in Chapter 3.3.2, can efficiently quantify uncertainty using a single deep neural network and a single forward pass, by employing a Dirichlet distribution in the output layer (Malinin and Gales, 2018; Sensoy et al., 2018). This efficiency advantage motivates us to explore EDL for mobile health diagnostics.

While EDL has shown great promise, most investigations and evaluations rely on well-curated datasets and balanced machine learning benchmarks such as CIFAR10 and CIFAR100 (Kopetzki

et al., 2021; Shen et al., 2023; Postels et al., 2022). However, we find that EDL can be ineffective due to the challenges posed by class imbalance, which is common in physiological data (as introduced in Chapter 1.2). The problem primarily stems from the uniform empirical loss across all samples and the assumption of a uniform distribution across all classes, as explained in detail in Chapter 5.3.3.

To make the efficient uncertainty quantification provided by EDL effective on multi-class imbalanced physiological data, this chapter proposes a novel class-balanced EDL approach. It introduces two mechanisms to enhance vanilla EDL: *i)* a class-level pooling loss to mitigate the bias in classification evidence and *ii)* a learnable prior that is regularised by the class distribution to facilitate learning for minority classes.

We demonstrate the superiority of our class-balanced EDL approach through comprehensive experiments conducted on a variety of physiological datasets and artificially imbalanced machine learning benchmark data. Our method not only reduces classification bias but also improves model calibration. Furthermore, we show that our uncertainty quantification is *accurate*, as it can be used to identify mistakes in model predictions and detect data that lie beyond the scope of the model's training, and also *efficient*, incurring no additional memory and computational costs compared to a single deep learning model. This accurate and efficient uncertainty quantification capability strengthens risk management and facilitates timely clinician involvement, thereby minimising the potential for misdiagnoses in automated mobile health applications.

The main contributions of this chapter are summarised below,

- We explore EDL for health diagnostics and introduce a novel class-balanced EDL approach to tackle the class imbalance challenge. This renders the efficient uncertainty-aware deep learning method effective even in the presence of class imbalance.
- We conduct extensive experiments using various data. The results reveal that our method not only improves diagnostic accuracy but also reduces overconfident predictions by up to 43% compared to the state-of-the-art baselines, without requiring additional costs to estimate the uncertainty.
- We introduce the use of uncertainty measurements for misdiagnosis identification and out-of-training-distribution detection. As a result, our method outperforms the compared baselines in these applications by up to 16.1% in terms of ROC-AUC.

The remainder of this chapter is organised as follows. We first review the related studies in Chapter 5.2. Then, we introduce our class-imbalanced EDL approach in Chapter 5.3. The experimental setup and results are presented from 5.4 to 5.6. We finally conclude our findings in Chapter 5.7.

5.2 Related work

Chapter 3.3.2 introduced the most representative uncertainty quantification methods and Chapter 4.2 discussed the application of model uncertainty for healthcare applications. In this section, we elaborate on the recent studies for EDL. Moreover, we extend the introduction to long-tailed learning in Chapter 3.3.1 to provide further justification for why existing methods are not applicable for EDL.

The concept of evidential deep learning was initially introduced in (Sensoy et al., 2018), where the Dirichlet distribution was proposed to address the issue of overconfidence stemming from the *Softmax* activation. Since then, various improvements have been proposed (Charpentier et al., 2020; Kopetzki et al., 2021; Ulmer, 2021). Notably, Charpentier *et al.* introduced a second-order uncertainty-aware loss function to enhance the learning of the Dirichlet distribution (Charpentier et al., 2020). This method is also the foundation of our study. In empirical demonstrations, Kopetzki *et al.* established that EDL stands as the new state-of-the-art uncertainty quantification method, proving competitive with supervised methods in terms of out-of-distribution detection (Kopetzki et al., 2021).

Regarding healthcare applications, the exploration of EDL has been limited. A literature review revealed that Li *et al.* proposed a region-based EDL segmentation framework capable of generating reliable uncertainty maps and accurate segmentation results. The results showcased the superior performance of the proposed method in quantifying segmentation uncertainty and robustly segmenting brain tumors (Li et al., 2023). To the best of our knowledge, we are the first to introduce EDL to applications in health diagnostics based on physiological data.

We also observe that a significant portion of EDL studies relies on meticulously curated datasets and balanced machine learning benchmarks, such as CIFAR10 and CIFAR100 (Kopetzki et al., 2021; Shen et al., 2023; Postels et al., 2022), leaving the implications of class-imbalanced data unclear. Despite the existence of various long-tailed learning approaches, encompassing both *data-level* and *algorithm-level* methods aimed at mitigating the adverse effects of class imbalance (as introduced in Chapter 3.3.1), these approaches are tailored for standard *Softmax*-based neural networks. As a result, most of them either cannot quantify model uncertainty or cannot be seamlessly integrated into the EDL framework, as EDL employs a different optimisation objective. For instance, focal loss is known for its effectiveness in addressing class imbalance by introducing a modulating factor called the focusing parameter, which reduces the loss for well-classified examples (Lin et al., 2017). However, this factor was designed for cross-entropy loss and thus cannot be applied to the Dirichlet distribution-based loss of EDL. In this chapter, we introduce novel components and learning strategies specifically designed to enhance EDL performance in the context of class-imbalanced data.

5.3 Methodology

5.3.1 Problem formulation

In this section, we introduce our class-balanced EDL approach to develop reliable health diagnostics models using imbalanced physiological data. For the sake of clarity, we first formulate the problem below.

Uncertainty-aware deep learning for multi-class health diagnostics: Consider a physiological dataset $\mathcal{D}$ with C categories (as defined in Chapter 3.1.1). Let N_c represent the number of samples for class c and N_c varies among classes. The task is to learn a deep neural network parameterised by θ that predicts y for any given x with an uncertainty measurement.

Before delving into the details of our method, we review the fundamentals of EDL to provide a comprehensive understanding. Subsequently, having identified the issues caused by class imbalance in EDL, we present our specific solutions in this section.

5.3.2 A recap for evidential deep learning

As formulated in Chapter 3.3.2, EDL leverages Dirichlet distribution $\boldsymbol{q}^{(i)}$ – the distribution over the categorical probability $\boldsymbol{p}^{(i)}$, to achieve prediction and uncertainty quantification simultaneously (Hastie et al., 2009; Murphy, 2012). As shown in Figure 3.5, the learnt posterior distribution $\boldsymbol{q}^{(i)} = Dir(\boldsymbol{\alpha}^{(i)})$ is parameterised by $\boldsymbol{\alpha}^{(i)} = [\alpha_1^{(i)}, \alpha_2^{(i)}, ..., \alpha_C^{(i)}]$ for C classes, where $\alpha_c^{(i)} = 1 + l_c^{(i)}$. For ease of understanding, the posterior Dirichlet distribution can be viewed as an infinite ensemble of point estimations $\boldsymbol{p}^{(i)}$. Therefore, EDL enables a better-calibrated way of quantifying *epistemic uncertainty* (as introduced in Chapter 3.3.2) compared to traditional *Softmax*-based deep learning (Malinin and Gales, 2018; Sensoy et al., 2018).

Additionally, for input $x^{(i)}$, the expectation of probability $\hat{\boldsymbol{p}}^{(i)}$ presents the average predictive confidence which reflects the *aleatoric uncertainty* (refer to Chapter 3.3.2). EDL is also able to capture the *distributional shift*: if no remarkable evidence can be modelled for a given input, the posterior $\alpha_c, \forall c \in C$ will approach 1, i,e., the prior. Overall, given an input $X^{(i)}$, an EDL model f_θ outputs distribution $\boldsymbol{q^{(i)}} = Dir(\boldsymbol{\alpha}^{(i)})$ with the predictive probability $\hat{\boldsymbol{p}}^{(i)}$, categorical prediction $\hat{y}^{(i)}$ inferred as below,

$$
\begin{aligned}
&\boldsymbol{\alpha}^{(i)} = \mathbf{1} + \boldsymbol{l}^{(i)}, \\
&\hat{p}^{(i)}[c] = \mathbb{E}[p^{(i)}[c]] = \frac{\alpha_c^{(i)}}{\alpha_0^{(i)}}, \\
&\hat{y}^{(i)} = \arg\max_c \mathbb{E}[p^{(i)}[c]],
\end{aligned}
\tag{5.1}
$$

where $\alpha_0^{(i)} = \sum_{c=1}^{C} \alpha_c^{(i)}$.

A few examples of Dirichlet distribution are illustrated in Figure 3.5 in Chapter 3.3.2. We hope to quantify both epistemic and aleatoric uncertainty and thus we adapted the following *Differential Entropy* ($DE^{(i)}$) as the measurement for predictive uncertainty,

$$DE^{(i)} = \mathbb{E}_{\boldsymbol{p}^{(i)} \sim \boldsymbol{q}^{(i)}}[Entropy(\boldsymbol{p}^{(i)})], \tag{5.2}$$

where $Entropy(\boldsymbol{p}^{(i)})$ captures the energy distributed across different classes (*i.e.*, aleatoric uncertainty) and the expectation reflect the "peakedness" in the Dirichlet distribution (*i.e.*, epistemic uncertainty). A larger DE corresponds to an overall higher uncertainty of a prediction.

Learning Objective. EDL can be adapted to any neural network architecture by simply replacing the *Softmax* layer with a plunge-in Dirichlet distribution estimation layer on the output side. Let $\mathcal{L}^{(i)}$ denote the loss for the i-th sample, to optimise the model parameters θ by feeding the training data set $\mathcal{D}$ with N samples, the following objective has been used for EDL (Charpentier et al., 2020; Bengs et al., 2022),

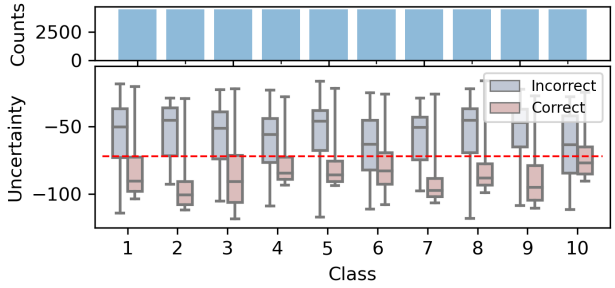
$$\begin{aligned} \min_{\theta} \ \mathcal{L} &= \frac{1}{N}\sum_{i=1}^{N} \mathcal{L}^{(i)}, \\ \mathcal{L}^{(i)} &= \mathbb{E}_{\boldsymbol{p}^{(i)} \sim \boldsymbol{q}^{(i)}}[\mathcal{C}(\boldsymbol{p}^{(i)}, y^{(i)})] + \lambda \cdot \mathcal{L}_r^{(i)}, \end{aligned} \tag{5.3}$$

where $\mathcal{C}$ denotes the cross-entropy, *i.e.*, $\mathcal{C}(\boldsymbol{p}^{(i)}, y^{(i)}) = -\log \boldsymbol{p}^{(i)}[y^{(i)}]$, $\mathcal{L}_r^{(i)}$ denotes a regularisation for each posterior $\boldsymbol{q}^{(i)}$. $\mathcal{L}^{(i)}$ is derived from variational inference: the optimisation of the posterior can be achieved by minimising the classification error and reducing the Kullback-Leibler divergence (KL divergence) between the posterior and prior (Cover, 1999). Specifically, the first term in $\mathcal{L}^{(i)}$ enforces the expected classification probability from the posterior Dirichlet distributions to be a good proxy of the ground-truth label. KL divergence is a measure from information theory that quantifies how much one probability distribution diverges from a second, expected probability distribution. Thus, the second term $\mathcal{L}_r^{(i)} = KL[(Dir(\boldsymbol{\alpha}^{(i)})||Dir(\mathbf{1})]$ enforces the the posterior to be similar with the prior distribution. λ represents the weight used to trade-off between the two terms. Finally, the total loss $\mathcal{L}$ is an empirical loss giving uniform importance to all training samples.

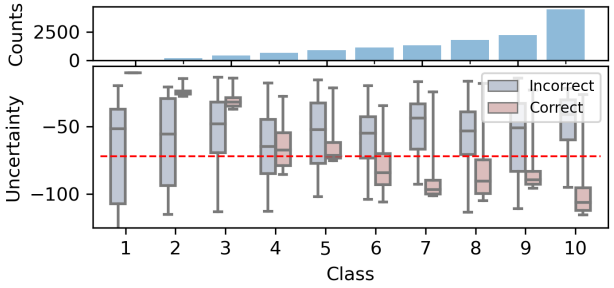
5.3.3 Impact of class imbalance on evidential deep learning

Now, we offer empirical and theoretical analyses of EDL to uncover its limitations in handling class-imbalanced data.

Empirical Observation. To demonstrate the degradation in performance of EDL when dealing with class imbalance, we conducted experiments using an artificially imbalanced dataset. Specifically, we implemented an image classification task using the EDL loss (Eq. (5.3)) and



(a) Balanced data yields reliable uncertainty estimates for all classes.



(b) Imbalanced training data leads to poor uncertainty estimates for the minority classes.

Figure 5.1: **Uncertainty quantified by EDL for CIFAR10 classification**. The top sub-figures present the training data distribution, and the bottoms show the uncertainty for correct and incorrect predictions within each class (a larger value indicates that the prediction is less certain). The red line represents an uncertainty threshold that leads to the highest accuracy in misclassification identification.

compared the results obtained with both balanced and down-sampled training data. After down-sampling, the data exhibited a skewed step distribution, mimicking class imbalance. For the experiments, the CIFAR-10 dataset, introduced in Chapter 2.2.3, is utilised. The model comprises VGG as the feature extractor and a classifier consisting of two fully connected layers (Tammina, 2019). The results of the experiments are presented in Figure 5.1, showcasing the distribution of the training data and the quality of the estimated uncertainty for each class.

As it can be observed in Figure 5.1(a), with balanced training data, for all 10 classes, the EDL quantifies higher uncertainty for incorrect predictions than correct predictions within each class. This suggests that quantified uncertainty can reliably reflect the confidence of the model, which is derived. However, this no longer holds with a skewed distribution as displayed in Figure 5.1(b): incorrect predictions from minority classes like class 0 and 1 manifest very low uncertainty. This evidence verifies our concern that the EDL is vulnerable in the class imbalance scenario, and modifications are necessary to improve its performance in healthcare applications.

Theoretical Analysis. In addition to empirical analysis, we also draw on theoretical insights to

systematically explain the reasons behind the failure of EDL in the presence of class imbalance.

Lemma I. *The across-sample empirical loss Eq. (5.3) induces the bias in EDL.*

Analysis. Given C classes, the objective in Eq. (5.3) can be rewritten as,

$$\begin{aligned}\min_{\theta} \ \mathcal{L} &= \frac{1}{N}\sum_{c=1}^{C}\sum_{y^{(i)}\in c}\mathcal{L}^{(i)} \\ &= \sum_{c=1}^{C}\frac{N_c}{N}\cdot\frac{1}{N_c}\sum_{y^{(i)}\in c}\mathcal{L}^{(i)} = \sum_{c=1}^{C}\frac{N_c}{N}\cdot\overline{\mathcal{L}_c},\end{aligned} \tag{5.4}$$

where $\overline{\mathcal{L}_c}$ presents the average loss for class c. It can be noted that class-averaged loss $\overline{\mathcal{L}_c}$ is weighted by the proportion of the samples in the training set. Herein, the object tends to prioritise optimising $\overline{\mathcal{L}_c}$ for the majority classes. Because of the relatively small N_c, misclassification or over-confident posteriors from minority classes could be under-looked, leading to imprecise estimation of classification evidence $\boldsymbol{l}$ (see Eq. (5.1)). Particularly, when N_c for minority classes is extremely small, which is common for many realistic applications where rare classes exist, the learned evidence can be more biased. As a consequence, the quantified uncertainty parameterised by $\boldsymbol{\alpha}$ could be less reliable for the minority classes due to the lack of training data.

Lemma II. *The uniform prior is not feasible for EDL in the presence of imbalanced data.*

Analysis. As EDL assumes a uniform $Dir(\mathbf{1})$ as a prior, it assumes an equal likelihood for all classes if the same amount of evidence has been observed. The regularisation of the posterior, *i.e.*, $\mathcal{L}_r^{(i)}$ in Eq. (5.3), also imposes a uniform smoothing across all classes, ignoring the varied learning difficulty among classes. This may not be optimal in the presence of imbalanced data, particularly when the minority classes are underrepresented with a few samples. In traditional *Softmax*-based deep learning, classification thresholds can be adjusted (*i.e.*, not the same threshold for every class) to allow some marginal samples to be classified into minority classes (Zou et al., 2016; Wang et al., 2019). Similarly, finding a suitable prior that can better regularise the posterior can be helpful in EDL.

5.3.4 Class-balanced evidential deep learning

Highlight*: Our method supposes that an adaptive prior and a loss function independent of class distribution can mitigate the bias introduced by the imbalanced physiological data in EDL, resulting in fairer and more reliable classification and uncertainty quantification for all classes.*

To address the challenges brought by class imbalance, we propose to optimise EDL by using a new objective to enable EDL for imbalanced physiological data.

New objective. Our efforts include two aspects: (1) learning less biased evidence and (2) seeking a better prior, which are introduced as follows,

Mechanism I. *Class pooling loss.* As discussed in *Lemma I*, the imbalanced distribution of samples among different classes acts as a significant source of bias in the model, resulting in varying learning rates across the classes. To overcome this issue, we propose to give equal attention to all classes no matter the number of training samples. To achieve this, we leverage a class-level pooling loss that is first calculated within each class and then averaged across classes. Specifically, $\mathcal{L}'$ in Eq. (5.3) will be replaced by,

$$\mathcal{L}' = \frac{1}{C}\sum_{c=1}^{C}\frac{1}{N_c}\sum_{y^{(i)}\in c}\mathcal{L}^{(i)}, \tag{5.5}$$

where N_c is the cardinality of class c. Thereby, in contrast to Eq. (5.4), $\mathcal{L}'$ is class distribution agnostic. In other words, we mitigate the bias by ensuring that the factor N_c/N approaches $1/C$ uniformly for all classes.

Mechanism II. *Adaptive prior.* Since the uniform prior assumption has limited capacity as discussed in *Lemma II.*, we propose to replace the uniform prior with a trainable prior parameterised by $\boldsymbol{\beta} = [\beta_1, \beta_2, ..., \beta_C]$. Learning the classification evidence through the neural network usually needs more data and could be biased, but optimising the posterior from the prior (*i.e.*, via $\mathcal{L}_r^{(i)}$) could be more helpful, particularly when the training data is limited. A good prior should consider the class distribution of the training data, and compensate for the class skew to ease the learning of the posterior. Herein, we propose that $\boldsymbol{\beta}$ can mimic the reversed class proportion, termed by $\boldsymbol{\eta} = [N/N_1, N/N_2, ...N/N_C]$ with $\eta_c = N/N_c$. Furthermore, although it is meaningful to use $\boldsymbol{\eta}$ as $\boldsymbol{\beta}$, we do not fix the value but use a trainable prior: this allows the prior with more optimisation space considering the varying learning difficulty for different classes. To achieve this, another term that measures the KL-divergence between the two categorical distributions parameterised by $\boldsymbol{\beta}$ and $\boldsymbol{\eta}$, formulated by,

$$\begin{aligned}\mathcal{L}'_p &= KL[Cat(\boldsymbol{\beta})||Cat(\boldsymbol{\eta})] \\ &= \sum_{c=1}^{C}\beta_c \log\frac{\beta_c}{\eta_c},\end{aligned} \tag{5.6}$$

will be added to the objective. Correspondingly, the regularisation term in Eq. 5.3 becomes the KL divergence between the posterior and the trainable prior to ensure "fidelity-to-prior" (Bengs

et al., 2022). We term the new posterior parameterised by $\boldsymbol{\alpha}'$. Following the definition of KL divergence in (Cover, 1999), we now provide the new regularisation for the posterior denoted by $\mathcal{L}_r'^{(i)}$ as follows,

$$\begin{aligned}\mathcal{L}_r'^{(i)} &= KL[Dir(\boldsymbol{\alpha}'^{(i)})||Dir(\boldsymbol{\beta})] \\ &= \int Dir(\boldsymbol{p}|\boldsymbol{\alpha}'^{(i)}) \log \frac{Dir(\boldsymbol{p}|\boldsymbol{\alpha}'^{(i)})}{Dir(\boldsymbol{p}|\boldsymbol{\beta})} d\boldsymbol{p} \\ &= \int Dir(\boldsymbol{p}|\boldsymbol{\alpha}'^{(i)}) \big(\log Dir(\boldsymbol{p}|\boldsymbol{\alpha}'^{(i)}) - \log Dir(\boldsymbol{p}|\boldsymbol{\beta}) \big) d\boldsymbol{p}.\end{aligned} \tag{5.7}$$

Since the integration can be derived by digamma function ψ and gamma function Γ, as,

$$\begin{aligned}&\int Dir(\boldsymbol{p}|\boldsymbol{\alpha}) \log Dir(\boldsymbol{p}|\boldsymbol{\alpha}) d\boldsymbol{p} \\ &= \int Dir(\boldsymbol{p}|\boldsymbol{\alpha}) \big[\log \Gamma(\alpha_0) - \sum_{c=1}^{C} \log \Gamma(\alpha_c) + \sum_{c=1}^{C} (\alpha_c - 1) \log \boldsymbol{p} \big] d\boldsymbol{p} \\ &= \log \Gamma(\alpha_0) - \sum_{c=1}^{C} \log \Gamma(\alpha_c) + \sum_{c=1}^{C} \alpha_c (\psi(\alpha_c) - \psi(\alpha_0)).\end{aligned} \tag{5.8}$$

The closed form of $\mathcal{L}_r'^{(i)}$ is written as,

$$\begin{aligned}\mathcal{L}_r'^{(i)} = \log \Gamma(\alpha_0'^{(i)}) - \sum_{c=1}^{C} \log \Gamma(\alpha_c'^{(i)}) - \log \Gamma(\beta_0) + \\ \sum_{c=1}^{C} \log \Gamma(\beta_c) + \sum_{c=1}^{C} (\alpha_c'^{(i)} - \beta_c)(\psi(\alpha_c'^{(i)}) - \psi(\alpha_0'^{(i)}))],\end{aligned} \tag{5.9}$$

where $\beta_0 = \sum_{c=1}^{C} \beta_c$.

Overall objective. Given the above, our proposed new optimisation loss for class-balanced EDL can be summarised as,

$$\begin{aligned}\min_{\theta,\boldsymbol{\beta}} \quad \mathcal{L}' &= \frac{1}{C} \sum_{c=1}^{C} \frac{1}{N_c} \sum_{y^{(i)} \in c} \mathcal{L}'^{(i)} + \mu \cdot \mathcal{L}_p', \\ \boldsymbol{\alpha}'^{(i)} &= \boldsymbol{\beta} + \boldsymbol{l}^{(i)}, \\ \mathcal{L}'^{(i)} &= \mathbb{E}_{\boldsymbol{p}^{(i)} \sim Dir(\boldsymbol{\alpha}'^{(i)})} [\mathcal{C}(\boldsymbol{p}^{(i)}, y^{(i)})] + \lambda \cdot \mathcal{L}_r'^{(i)}, \\ \mathcal{L}_r'^{(i)} &= KL[Dir(\boldsymbol{\alpha}'^{(i)})||Dir(\boldsymbol{\beta})], \\ \mathcal{L}_p' &= KL[Cat(\boldsymbol{\beta})||Cat(\boldsymbol{\eta})],\end{aligned} \tag{5.10}$$

where hyper-parameters λ and μ trade off the classification, the regularisation of posterior $\mathcal{L}_r'$,

and the regularisation of prior $\mathcal{L}'_p$.

Now, we give the closed form of the loss function and show that the model can be optimised without sampling from the Dirichlet distribution. Specifically, the closed form of the expected cross-entropy can be derived as,

$$
\begin{aligned}
&\mathbb{E}_{\boldsymbol{p}^{(i)}\sim Dir(\boldsymbol{\alpha}'^{(i)})}[\mathcal{C}(\boldsymbol{p}^{(i)}, y^{(i)})] \\
=&\mathbb{E}_{Dir(\boldsymbol{p}|\boldsymbol{\alpha}'^{(i)})}[-\log p_{y^{(i)}}] \\
=&-\int \log p_{y^{(i)}} \cdot Dir(\boldsymbol{p}|\boldsymbol{\alpha}'^{(i)}) d\boldsymbol{p} \\
=&-\int_0^1 \log p_{y^{(i)}} \cdot Beta(\alpha'^{(i)}_{y^{(i)}}, \alpha'^{(i)}_0 - \alpha'^{(i)}_{y^{(i)}}) dp_{y^{(i)}} \\
=&-\frac{\int_0^1 \frac{dp_{y^{(i)}}^{\alpha'^{(i)}_{y^{(i)}}-1}}{d\alpha'^{(i)}_{y^{(i)}}} (1-p_{y(i)})^{\alpha'^{(i)}_0 - \alpha'^{(i)}_{y^{(i)}} - 1} dp_{y^{(i)}}}{Beta(\alpha'^{(i)}_{y^{(i)}}, \alpha'^{(i)}_0 - \alpha'^{(i)}_{y^{(i)}})} \\
=&-\frac{1}{Beta(\alpha'^{(i)}_{y^{(i)}}, \alpha'^{(i)}_0 - \alpha'^{(i)}_{y^{(i)}})} \frac{dBeta(\alpha'^{(i)}_{y^{(i)}}, \alpha'^{(i)}_0 - \alpha'^{(i)}_{y^{(i)}})}{d\alpha'^{(i)}_{y^{(i)}}} \\
=&-\frac{d\log Beta(\alpha'^{(i)}_{y^{(i)}}, \alpha'^{(i)}_0 - \alpha'^{(i)}_{y^{(i)}})}{d\alpha'^{(i)}_{y^{(i)}}} \\
=&-\frac{d(\log\Gamma(\alpha'^{(i)}_{y^{(i)}}) + \log\Gamma(\alpha'^{(i)}_0 - \alpha'^{(i)}_{y^{(i)}}) - \log\Gamma(\alpha'^{(i)}_0))}{d\alpha'^{(i)}_{y^{(i)}}} \\
=&\frac{d\log\Gamma(\alpha'^{(i)}_0)}{d\alpha'^{(i)}_0} - \frac{d\log\Gamma(\alpha'^{(i)}_{y^{(i)}})}{d\alpha'^{(i)}_{y^{(i)}}} \\
=&\psi(\alpha'^{(i)}_0) - \psi(\alpha'^{(i)}_{y^{(i)}}),
\end{aligned}
\tag{5.11}
$$

where Γ denotes the gamma function, ψ is the digamma function, and $\alpha'^{(i)}_0 = \sum_{c=1}^{K} \alpha'^{(i)}_c$.

Integrating Eq. (5.6), (5.9), (5.11) into Eq. (5.10), we give the overall loss function as,

$$
\mathcal{L}' = \frac{1}{C}\sum_{c=1}^{C} \frac{1}{N_c} \sum_{y^{(i)}\in c} \{\mathbb{E}_{\boldsymbol{p}^{(i)}\sim\boldsymbol{q}^{(i)}}[\mathcal{C}(\boldsymbol{p}^{(i)}, y^{(i)})] + \lambda \cdot \mathcal{L}_r^{(i)}\} + \mu \cdot \sum_{c=1}^{C} \beta_c \log\frac{\beta_c}{\eta_c}, \tag{5.12}
$$

where $\mathbb{E}_{\boldsymbol{p}^{(i)}\sim\boldsymbol{q}^{(i)}}[\mathcal{C}(\boldsymbol{p}^{(i)}, y^{(i)})] = \psi(\alpha'^{(i)}_0) - \psi(\alpha'^{(i)}_{y^{(i)}})$, $\mathcal{L}_r'^{(i)} = \log\Gamma(\alpha'^{(i)}_0) - \sum_{c=1}^{C}\log\Gamma(\alpha'^{(i)}_c) - \log\Gamma(\beta_0) + \sum_{c=1}^{C}\log\Gamma(\beta_c) + \sum_{c=1}^{C}(\alpha'^{(i)}_c - \beta_c)(\psi(\alpha'^{(i)}_c) - \psi(\alpha'^{(i)}_0))]$, $\alpha'^{(i)}_0 = \sum_{c=1}^{C}\alpha'^{(i)}_c$, $\beta_0 = \sum_{c=1}^{C}\beta_c$, ψ is the digamma function and Γ denotes the gamma function. Now, we can effectively optimise the model via gradient descent and backpropagation, as introduced in Chapter 3.1.2.

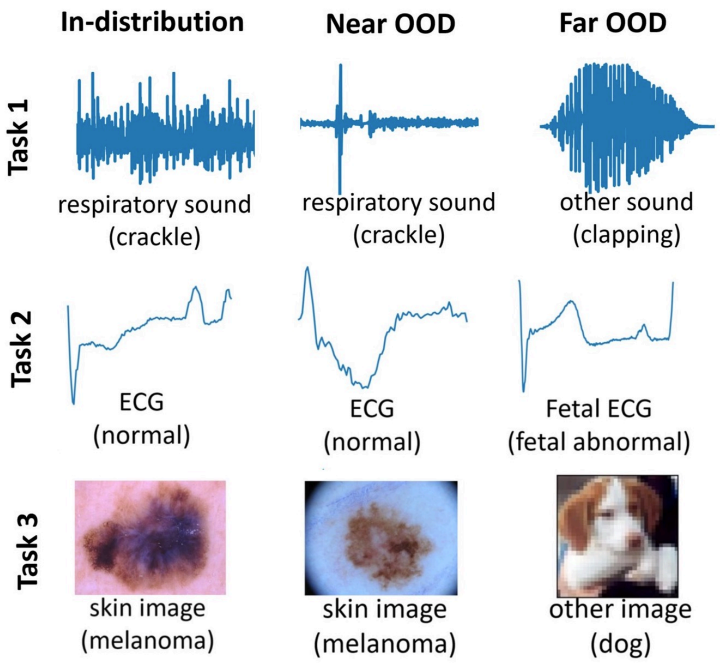


Figure 5.2: **Data examples**. In and out-of-distribution testing samples are given for the three tasks.

5.4 Experimental setup

To validate the effectiveness of our class-balanced EDL for real-world multi-class health diagnostics, we employ three clinical tasks for experiments. The datasets used encompass various physiological data modalities, and all of them exhibit severe class imbalance, making them ideal test beds for evaluation.

5.4.1 Datasets and task setup

We conduct extensive experiments on three clinical tasks with different physiological data modalities. We split each dataset into a training and a testing set. The training set including a part for validation is used for model parameter learning, while the testing set is leveraged to report the performance. For each task, we also include two OOD (out-of-training-distribution) testing sets, *i.e.*, near OOD and far OOD. The near OOD set has the same classes as the training data but was collected with a different protocol, thus presenting a semantic shift, while the far OOD set contains similar inputs to unseen classes. An overview of the used datasets and models are summarised in Table 5.1. These models comprise a feature extractor and a classifier consisting of two fully connected layers. Examples of the in and out-of-distribution samples are given in Figure 5.2. The details of those datasets are elaborated below.

Task 1: Respiratory abnormality detection. We explore the potential of lung sounds for de-

tecting respiratory abnormalities by distinguishing abnormal lung sounds from healthy sounds, leveraging the state-of-the-art ResNet34-based acoustic model (Gairola et al., 2021). The model used for this task is similar to the audio-driven COVID-19 screening model introduced in Chapter 4. The lung sound samples are converted into spectrograms, from which deep neural networks extract features for classification.

- **(ID)** We use the ICBHI 2017 Respiratory Challenge data for training and the in-distribution (ID) testing set (Rocha et al., 2019) (details can be found in Chapter 2.2.1). The total 6,898 samples from 126 patients cover four classes: normal lung sounds (52.8%), crackle only (27.0%), wheeze only (12.9%), and both crackle and wheeze (7.3%).
- **(Near OOD)** A similar audio dataset named Stethoscope consists of 336 normal, crackle, and wheeze audio samples (Fraiwan et al., 2021). we use it as ICBHI's co-variate shift counterpart, as although this dataset covers the same pathology, it is collected from a 3M Littemann electronic Stethoscope, differing from ICBHI.
- **(Far OOD)** ARCA23K is a dataset of labelled sound events originating from *Freesound*, and each clip belongs to one of 70 typically audio classes including music, human sounds, animal sounds, etc.[1]. We used the validation set containing 2,264 clips.

Setting. For the ID data, we follow the official patient-independent training and testing splits of the Challenge. Samples from 47 patients are used for testing, while for the rest of the patients, we randomly divide them into five folds and hold out one fold per run to conduct five-fold cross-validation. For all ID and OOD datasets, audio recordings are re-sampled to 4KHz and divided into 8s clips. The clips are then transformed into Mel-spectrograms as the inputs of the model.

Task 2: Heart failure prediction. The detection of cardiovascular diseases is investigated using electrocardiogram (ECG) data with the one-dimensional convolutional neural network FCNet (Avanzato and Beritelli, 2020).

- **(ID)** ECG5000 is a 20-hour long one-channel ECG dataset, which has been split and interpolated into equal-length (140) heartbeats. We use this database as the ID set (refer to Chapter 2.2.2). It consists of five classes: 58.4% are normal, 35.3% have heat failure typed R-on-T, 3.9% PVC, 2.0% SP, and 0.5% UB.
- **(Near OOD)** Another dataset consisting of 200 ECG recordings with a length of 178 is used as the near OOD[2], because the data acquisition method is different from ECG5000 (Olszewski, 2001).
- **(Far OOD)** A non-invasive fetal ECG dataset consists of 1,965 heartbeats with a length

[1] https://zenodo.org/record/5117901#.YkCsRk3MJPY
[2] https://timeseriesclassification.com/description.php?Dataset=ECG200

Table 5.1: **A summary of the used physiological datasets**. #Train is the original training data size, which is split into training and validation folds with different seeds. #Test is the testing size. C is the number of classes.

Task			ID Dataset					OOD Dataset			
Name	Extractor	Modality	Name	#Train	#Test	C	Ratio (%)	Near OOD	Size	Far OOD	Size
Task 1	ResNet34	Audio	ICBHI2017	4,274	2,641	4	52.8/27.0/12.9/7.3	Stethoscope	336	ARCA23K	2,264
Task 2	FCNet	ECG	ECG5000	4,500	500	5	58.4/35.3/3.9/2.0/0.5	ECG200	200	FetalECG	1,965
Task 3	DenseNet121	Image	HAM10000	7,206	2,809	7	67.1/11.1/11.0/5.1/3.3/1.4/1.1	ISIC2017	1,824	CIFAR-10	10,000

of 750[3]. As electrodes were placed on the mother's abdomen, the ECG is usually of lower amplitude than the maternal's, and thus we use it as the far ODD dataset.

Setting. We utilise a subset of 500 samples in the ID ECG5000 datasets for testing and split the rest into five folds uniformly for cross-validation.

Task 3: Skin lesion screening. The classification of skin lesions is examined using an image classification model based on DenseNet121 (Pacheco et al., 2020).

- **(ID)** We leverage the HAM10000 (Tschandl et al., 2018) for training and the ID testing set (refer to Chapter 2.2.3). The skin condition is labelled as one of the following classes: melanocytic nevi (67.1%), melanoma (11.1%), benign keratosis-like lesion (11.0%), basal cell carcinoma (5.1%), actinic keratoses (3.3%), vascular lesion (1.4%), or dermatofibroma (1.1%).
- **(Near OOD)** Another skin lesion dataset with 2,000 high-resolution varied-size images published by ISIC 2017 is used (Codella et al., 2018). It was collected by another institute with a varied device from HAM10000, therefore we regard it as the near OOD.
- **(Far OOD)** The image classification benchmark CIFAR10 (refer to Chapter 2.2.3) with 10 non-skin classes is utilised as the far OOD.

Setting. For ID data, 30% is held out as the testing set, and five-fold cross-validation is implemented: four-fifths of the remaining 70% of the data for training and one-fifth for validation per running. Image augmentation is conducted by slightly modifying the brightness of the images in the minority classes for training.

Machine learning benchmark data study. In addition to the above physiological datasets, we also include the machine learning benchmark data to evaluate the general capability of our method. Specifically, we use CIFAR10 as introduced in Chapter 2.2.3 for the training and in-distributional testing set. We employ the Street View House Numbers (SVHN) dataset[4] (con-

[3] https://timeseriesclassification.com/description.php?Dataset=NonInvasiveFetalECGThorax1

[4] http://ufldl.stanford.edu/housenumbers/

taining different image classes from CIFAR10) as an out-of-distribution testing set.

Setting. Being consistent with the empirical study in Chapter 5.3.3, VGG is used as the feature extractor. To provide a more quantitative analysis, we generate different levels of class imbalance: we down-sample the original CIFAR10 training set while preserving a uniform distribution among classes. We term the imbalance ratio as the ratio between the size of the largest category and the smallest class (Huang et al., 2022b). We create light (ratio=10), mild (ratio=50), and heavy (ratio=100) imbalances for training.

5.4.2 Baselines and metrics

The model with *Softmax* probability, termed as **Vanilla**, is implemented for each task as a basic baseline. Besides, we compare our method to the state-of-the-art long-tailed learning methods and uncertainty estimation methods, respectively.

For the former group, we include typical re-balancing approaches: weighted cross-entropy loss (**WL**) (Aurelio et al., 2019) and random-over-sampling (**ROS**) (Shelke et al., 2017). We also employ a recently proposed supervised deep clustering method (**SDC**) (Öztürk and Çukur, 2022). SDC first learns the class embeddings by maximising cluster separation and then uses a novel triplet loss to discriminate the learned embeddings. This two-stage learning protocol improves the reliability against imbalanced training data.

For the latter group, we first report the performance of EDL optimised by Eq. (5.3), which is termed as **Vanilla EDL** without re-balancing the class. We also compare EDL with the other two uncertainty quantification approaches. The first approach is the Monte Carlo Dropout method (referred to as **MCDrop**) (Gal and Ghahramani, 2016a; Lemay et al., 2022), which captures model uncertainty by keeping dropout activated during testing. The other approach is deep ensemble learning (referred to as **Ensemble**), which quantifies uncertainty based on the outputs of multiple models (Lakshminarayanan et al., 2017; Xia et al., 2021a). Although these methods have shown promise in well-curated data, they were not specifically designed for imbalanced data. To ensure a fair comparison, we implemented them using the same data augmentation techniques as EDL, namely **MCDrop+ROS** and **Ensemble+ROS**. It can be noted that **Ensemble+ROS** is an extension of the method we developed in Chapter 4. For **MCDrop**, during inference, we use a Dropout rate of 50% and run the model five times. For **Ensemble**, we train five model units using different random seeds for data augmentation. For these two methods, the final prediction of each instance is derived through probability-wise fusion, as introduced in Chapter 4.3.2. Instead of using the DE metric, the uncertainty measurement for non-EDL methods was the entropy of the predictive probabilities (Qendro et al., 2021a).

For all the methods in this section, we use a learning rate of 10^{-4}, the Adam optimiser, a batch

size of 64 (unless specifically mentioned), and a maximum epoch of 200. The best model based on the highest accuracy on the validation set is saved. ResNet-34 and DenseNet-121 are pre-trained by image data benchmark, while other parameters are randomly initialised.

For evaluation purpose, we report accuracy-centric metric ***Rec*** and uncertainty-centric metrics ***Brier***, ***ECE***. *Rec* is the macro-recall (macro-sensitivity) on the testing set, denoted by, $Rec = \frac{1}{C}\sum_{c=1}^{C} ACC(\hat{y}^{(i)}|y^{(i)} = c)$ ($\hat{y}^{(i)}$ is the prediction and $y^{(i)}$ is the ground truth). *Rec* evaluates the overall accuracy of categorical predictions, while *Brier* and *ECE* assess the calibration of predicted probabilities (Postels et al., 2022). A detailed formulation of them can be found in Chapter 3.4. For *ECE*, we partition the estimated confidence into $M = 10$ equal bins on the testing set. For a calibrated model, it is desired to minimise the values of *Brier* and *ECE*, while maintaining the same or higher values for *Rec*.

Additional, we present two uncertainty measurement-driven applications: ***misclassification identification*** and ***out-of-distribution (OOD) detection*** (Shen et al., 2023). We evaluate the performance by $\boldsymbol{AUC}_m$ and $\boldsymbol{AUC}_o$ for the two tasks, respectively. AUC, short for ROC-AUC (refer to Chapter 3.4), is used to measure the accuracy of classification. We treat the evaluation as a binary classification task: misclassified/OOD data belongs to the positive class while correctly predicted/ID data is the negative class. We conduct min-max normalisation for uncertainty measurements on the testing set (for EDL methods, we use DE, and for other baselines, we use $Entropy$), resulting in the normalised values ranging $[0, 1]$. Those normalised uncertainty measurements are the probabilities to calculate AUC. To distinguish between near and far out-of-distribution (OOD) detection, AUC_o^n and AUC_o^f are reported, respectively. A higher value for AUC_m and AUC_o indicates better utilisation of the quantified uncertainty measurements to ensure the safety of health diagnostics provided by the model.

5.5 Results on various imbalanced physiological data

5.5.1 Overall performance comparison

For the three clinical tasks involving imbalanced physiological data in model development, the results are summarised in Table 5.2 and will be discussed below.

Task 1. The task involves a 4-class classification problem with mildly imbalanced data (refer to Table 5.1). The first observation is that both Vanilla and Vanilla EDL struggle to perform well, while the re-balancing strategy WL and ROS significantly improve the Vanilla and Vanilla EDL across all the metrics. SDC is a strong baseline for class imbalanced data by ensuring the class margin, but it is still a deterministic model using Softmax to generate the final prediction, which indicates that the model could be over-confident for out-of-distribution data. As proven by the

Table 5.2: **Performance comparison**. The average results of five runs are shown. The best results are highlighted and the second best are underlined for each metric.

	Rec↑	Brier↓	ECE↓	AUC_m↑	AUC_o^n↑	AUC_o^f↑
	Task 1: Respiratory abnormality detection					
Vanilla	0.256	0.999	0.310	0.587	0.650	0.728
WL	0.401	0.949	0.292	0.594	0.661	0.698
ROS	0.407	0.941	0.301	0.605	0.673	0.742
SDC	0.422	0.902	0.288	0.617	0.664	0.747
Vanilla EDL	0.268	0.983	0.304	0.603	0.655	0.734
EDL+WL	0.389	0.908	0.290	0.621	0.687	0.759
EDL+ROS	0.434	0.878	0.297	0.620	0.700	0.768
MCDrop+ROS	0.412	0.933	0.289	0.625	0.690	0.764
Ensemble+ROS	0.431	0.929	0.286	0.628	0.699	0.769
Ours	0.422	0.797	0.163	0.640	0.727	0.785
	Task 2: Heart failure prediction					
Vanilla	0.389	0.690	0.179	0.850	0.782	0.885
WL	0.715	0.480	0.073	0.608	0.690	0.766
ROS	0.717	0.482	0.071	0.597	0.681	0.758
SDC	0.732	0.476	0.073	0.600	0.692	0.770
Vanilla EDL	0.388	0.685	0.175	0.843	0.786	0.887
EDL+WL	0.585	0.521	0.123	0.622	0.788	0.893
EDL+ROS	0.690	0.478	0.062	0.848	0.790	0.920
MCDrop+ROS	0.721	0.471	0.067	0.602	0.707	0.772
Ensemble+ROS	0.728	0.452	0.068	0.598	0.708	0.798
Ours	0.778	0.319	0.062	0.911	0.917	0.973
	Task 3: Skin lesion screening					
Vanilla	0.610	0.538	0.217	0.740	0.695	0.789
WL	0.689	0.457	0.159	0.784	0.665	0.891
ROS	0.727	0.441	0.110	0.801	0.693	0.927
SDC	0.730	0.439	0.112	0.813	0.705	0.927
Vanilla EDL	0.601	0.534	0.214	0.747	0.688	0.803
EDL+WL	0.678	0.511	0.153	0.798	0.694	0.882
EDL+ROS	0.735	0.428	0.105	0.830	0.701	0.896
MCDrop+ROS	0.734	0.429	0.103	0.835	0.735	0.949
Ensemble+ROS	0.739	0.420	0.102	0.840	0.735	0.950
Ours	0.763	0.396	0.095	0.854	0.747	0.968

results, the uncertainty-aware baselines, *i.e.*, EDL+WL, EDL+ROS, MCDrop+ROS, and Ensemble+ROS, generally perform better for uncertainty-centric metrics. However, within those methods, none of them consistently outperforms the others across all metrics, highlighting the challenge of achieving accurate diagnosis accuracy and high-quality uncertainty measurements simultaneously in real-world applications. We recognise that this difficulty primarily stems from the heterogeneity of the data, as the audio recordings were collected using different stethoscopes. Thus, an effective uncertainty estimation method is necessary to accurately quantify the uncertainty from both the data and the model.

In comparison to the baselines, our class-balanced EDL approach achieves competitive results in terms of *Rec*. Although a *Rec* of 0.422 is not the best, it is very close to the best of 0.434.

Yet, our method demonstrates significantly superior uncertainty measurements. Notably, we have successfully reduced *ECE* by 43% (from 0.286 to 0.163), indicating that our model can effectively avoid overconfident detection of respiratory abnormalities. The accuracy of detecting misclassification and OOD is also improved by 2.4%, 3.9% and 2.1% compared to the second best as underlined in Table 5.2, respectively.

Task 2. In this task, the physiological data is highly imbalanced, with the three minority classes accounting for less than 10% of the data. From Table 5.2, it can be observed that with such severe class imbalance, SDC achieves the highest *Rec* of 0.732 among the compared methods. Baselines including EDL+WL, EDL+ROS, MCDrop+ROS, and Ensemble+ROS significantly improve the classification performance as measured by *Rec*, and reduce overconfident predictions as reducing *Brier* and *ECE*. However, they fail to improve the utilisation of uncertainty measurements, *i.e.*, no better *AUC*s. It is plausible that the baselines with weighted loss or data augmentation mechanisms can effectively reduce bias in classification, but they are unable to mitigate bias in uncertainty quantification.

Obtaining accurate uncertainty estimation for this task is particularly challenging compared to the other two tasks. Task 2 presents several difficulties due to its smaller training dataset and the close semantic similarity between the OOD data and the ID ECG data. This challenge becomes evident when we observe that none of the baseline methods consistently outperforms the others across all metrics. Notably, it leads to a remarkable increase of 6.9% in *Rec*, a substantial reduction of 29.4% in *Brier*, and a significant enhancement in the *AUC* of detecting misclassification and OOD samples, with improvements ranging from 5.8% to 16.1%, respectively.

Task 3. In Task 3, the training data consists of 67.1% images from healthy subjects, while the remaining data comprises six other types of lesions, exhibiting a long-tailed distribution. On this type of data, vanilla methods (Vanilla and Vanilla EDL) are vulnerable and all other methods outperform them in terms of classification and uncertainty quantification.

Among the baselines, Ensemble+ROS achieves the best performance. However, our class-balanced EDL still exhibits performance gains compared to Ensemble+ROS for all of the metrics. Specifically, we can observe the improvements of 3.2% in balanced *Rec*, 5.7% in *Brier*, 5.7% in *ECE*, and about 2% in AUC_m, AUC_o^n, and AUC_o^f. It is also worth mentioning that the Ensemble baseline requires multiple passes during inference, making it less efficient compared to our method. These observations empirically validate the superiority of our methods over the compared baselines.

To summarise the above results, our method, which encompasses the joint optimisation of EDL posterior and prior, not only enhances classification performance but also concurrently improves the quality of uncertainty estimation for health diagnostics. This suggests the effectiveness of

Table 5.3: **Results for memory and computational costs**. Size: number of model parameters. FLOPS: number of floating point operation per instance during inference.

	Task 1: ResNet34		**Task 2: FCNet**		**Task 3: DenseNet121**	
	Size ($\times 1e^6$)	FLOPs ($\times 1e^9$)	Size ($\times 1e^6$)	FLOPs ($\times 1e^6$)	Size ($\times 1e^6$)	FLOPs ($\times 1e^9$)
Vanilla	21.39	10.75	0.19	17.26	7.13	45.43
MCDrop	21.39	53.70	0.19	86.24	7.13	227.14
Ensemble	106.95	53.75	0.98	86.24	35.65	227.15
Ours	21.39	10.75	0.19	17.26	7.13	45.43

our mechanisms in addressing challenges arising from imbalanced physiological data.

5.5.2 Efficiency Analysis

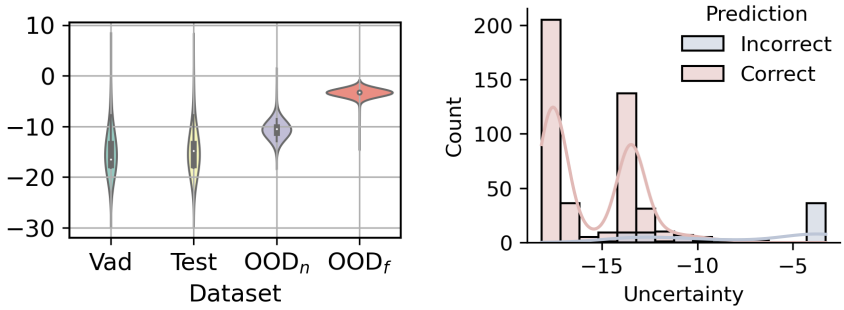
Results in Table 5.2 provide compelling evidence that our class-balanced EDL method outperforms the compared methods in both classification accuracy and the quality of uncertainty estimates. In addition to these achievements, we delve into an analysis of our method's superiority in terms of efficiency.

To this end, we compare the memory and computational costs during inference required by our method against other baselines: Vanilla, MCDrop, and Ensemble. The results are succinctly summarised in Table 5.3. Memory cost is measured by the model size, equivalent to the number of parameters stored in memory, while the count of floating-point operations (FLOPs) is used to assess computational cost.

Table 5.3 reveals that our EDL method incurs the same memory and computational costs as the Vanilla method, *i.e.*, a single model with *Softmax* output. However, our model's performance significantly surpasses that of Vanilla, as demonstrated in Table 5.2. Notably, our approach saves five times the FLOPs compared to MCDrop, as MCDrop requires running the model five times to estimate uncertainty. While the Ensemble approach, with data augmentation for training, performs well for the three tasks summarised in Table 5.2, its memory and computation costs are five times higher than our method. This validates our motivation that despite the ensemble learning approach's excellent performance in most cases, its inefficiency makes it challenging to deploy it for resource-constrained uncertainty quantification. These results collectively demonstrate the superior efficiency of our method which does not incur additional memory and computation costs compared to traditional *Softmax*-based deep learning.

5.5.3 Implications of uncertainty quantification

To gain a deeper understanding, we visualise the uncertainty distribution for the training, validation, and testing sets of Task 2 in Figure 5.3(a). It is evident that the near and far OOD sets



(a) DE for all testing sets (Vad is the validation set).

(b) ID testing set.

Figure 5.3: **Uncertainty distribution**. The uncertainty is measured by DE for heart failure prediction (Task 2).

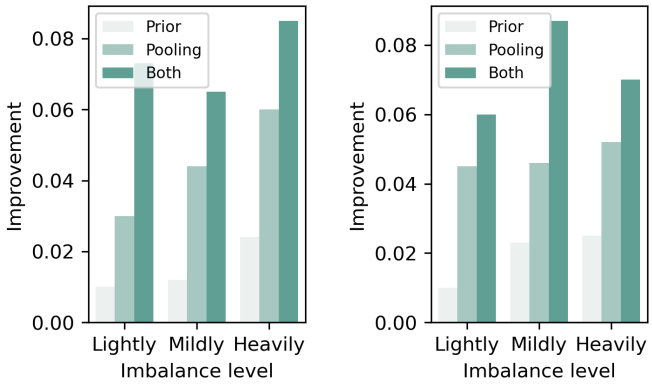
exhibit larger uncertainty measurements compared to the validation and ID testing sets, with the far OOD set displaying even greater uncertainty. This observation implies that an uncertainty threshold can be identified from the validation set and utilised to reject certain automated diagnoses made by the system. This approach effectively reduces the risk of misdiagnosis caused by shifts in the data distribution.

Within the ID testing set, we further divide the predictions into correct and incorrect prediction groups, and their corresponding uncertainties are displayed in Figure 5.3(b). It is clearly observed that correct predictions tend to have lower uncertainty compared to incorrect predictions. This suggests that, even for in-distributional data, the model may fail to diagnose certain challenging cases. However, our method is able to generate high uncertainty for those misdiagnoses, thereby improving the reliability of the system.

In conclusion, our proposed class-balanced EDL method proves highly accurate for diagnostics and uncertainty quantification in various imbalanced physiological data scenarios. It not only demonstrates significant improvements over vanilla EDL but also outperforms the compared baseline methods, especially in cases of extreme data imbalance. Notably, our method can generate high-quality uncertainty estimates without incurring additional computing costs compared to the traditional *Softmax*-based deep learning approach. Therefore, it can meet the efficiency requirements of mobile health applications. These results pave the way for deploying reliable deep learning-driven health diagnostics using the physiological data collected from mobile devices in real-world settings.

Table 5.4: **Performance by vanilla EDL and our class-balanced EDL on CIFAR10 with various imbalance levels**. The arrows after the metrics indicate the optimal direction. Mean±std across five runs is reported. The best results are highlighted.

		Rec↑	Brier↓	ECE↓	AUC_m↑	AUC_o↑
Balanced	Vanilla	0.871±0.003	0.219±0.016	0.100±0.008	0.815±0.015	0.801±0.008
Lightly Imbalanced	Vanilla	0.830±0.008	0.348±0.032	0.134±0.014	0.723±0.018	0.780±0.010
	Ours	0.833±0.005	0.325±0.024	0.120±0.009	0.796±0.017	0.786±0.010
Mildly Imbalanced	Vanilla	0.764±0.009	0.434±0.016	0.166±0.014	0.688±0.033	0.650±0.057
	Ours	0.781±0.008	0.389±0.014	0.136±0.013	0.753±0.034	0.737±0.048
Heavily Imbalanced	Vanilla	0.700±0.025	0.550±0.047	0.213±0.022	0.643±0.075	0.627±0.084
	Ours	0.734±0.025	0.402±0.039	0.157±0.021	0.728±0.078	0.697±0.082



(a) For misclassification. (b) For OOD detection.

Figure 5.4: **Result comparison for ablation study**. Absolute improvement of *AUC* from vanilla EDL on two uncertainty-driven applications are visualised.

5.6 Results on machine learning benchmark data

In the previous section, we showcased the effectiveness of our proposed method on three real-world physiological datasets. To reinforce the breadth of our evaluation, in this section, we present extended experimental results on CIFAR10.

Overall comparison. The performance comparison is summarised in Table 5.4. The balanced group shows the results of EDL on the original class-balanced CIFAR10 training data, outperforming all other groups and serving as an upper bound. Across lightly to heavily imbalanced scenarios, both vanilla EDL and our proposed mechanism exhibit a decline in performance, with vanilla EDL experiencing a decrease ranging from 2.6% to 151.1%, and our mechanism showing a decrease from 1.9% to 83.6% across all metrics. This suggests that class imbalance poses a great challenge to EDL yet our proposed mechanism performs better compared to vanilla EDL in varying degrees of class imbalance.

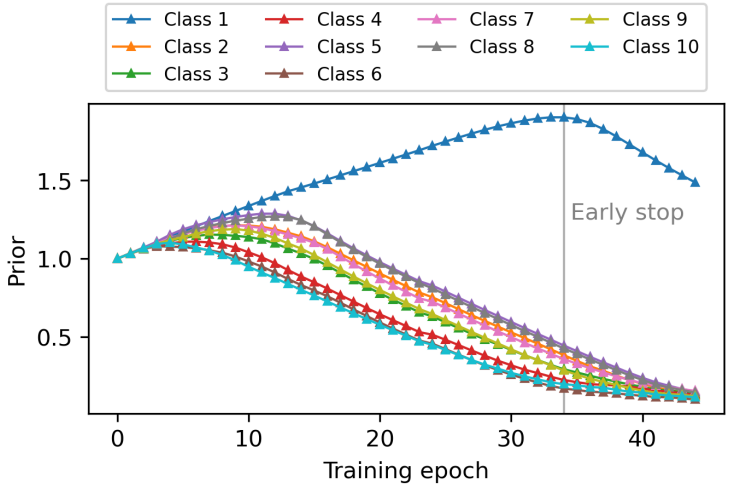


Figure 5.5: **Optimising the parameters of the prior**. The plot presents the updates of the prior β_c in a heavily imbalanced example, where class 10 has the largest training set and class 1 has the smallest data proportion (refer to Figure 5.1(b)).

Notably, the performance improvement over vanilla EDL is particularly prominent in heavily imbalanced cases. Our mechanism achieves a 4.9% increase in *Rec*, over a 26.1% reduction in *Brier* and *ECE*, and an improvement of more than 10% in AUC_m as well as AUC_o. These results confirm the effectiveness of our modifications to EDL in mitigating the bias caused by imbalanced training data, resulting in more accurate classification and reliable uncertainty estimation, especially for heavily imbalanced datasets.

Ablation study. We also investigate the individual contributions of each mechanism in our method. To do so, we conduct experiments on CIFAR10 using only a trainable prior with the loss specified in Eq. (5.10), or a pooling loss with a uniform prior. Independent improvements in each group can be observed, but the combination of the two mechanisms results in the most significant performance gain compared to vanilla EDL. Figure 5.4 illustrates the results. For the two uncertainty-driven applications, applying the trainable prior led to an improvement of 0.01 to 0.02 in AUC, while the deployment of the pooling loss increased AUC by 0.03 to 0.06. Interestingly, in most cases, the combined use of the two mechanisms yielded a greater improvement in performance compared to the sum of the improvements achieved separately. This can be attributed to the fact that the joint learning of the prior and posterior, regularised by the data distribution, creates a larger optimisation space for the model. Consequently, when the model converges, it achieves better performance due to this expanded optimisation space.

Additionally, we visualise the learning of the prior in Figure 5.5. The prior $\boldsymbol{\beta}$ for each class dynamically changes as the training progresses, initially increasing and then reaching a stable value. Notably, β_1 consistently remains larger than other β_c values, as the training data for class 1 is the smallest compared to the others. While learning the classification evidence through the

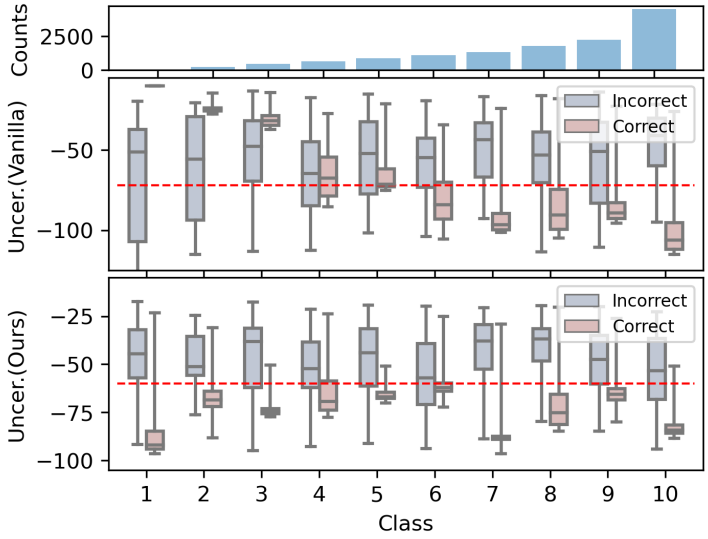
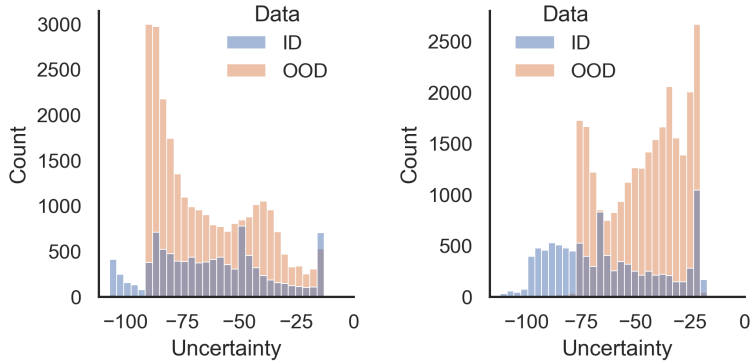


Figure 5.6: **Uncertainty performance in the application of misclassification detection**. For a heavily imbalanced case, our method significantly improves the quality of uncertainty. The results can also be compared with Figure 5.1.

neural network typically requires more data, directly learning the prior from the class distribution is less data-intensive. Through this approach, we effectively mitigate the bias in the posterior by leveraging an easily calibrated prior to compensate for the challenging-to-learn evidence. It is also interesting to observe that in this example, the best model is achieved when β_1 reaches its peak. This further validates our assumption that a well-regularised prior, guided by the class distribution, facilitates the learning of the posterior.

Implications of uncertainty quantification. To show that the quantified uncertainty measurements can benefit misclassification and OOD detection, we present the following case studies. Considering a heavily imbalanced scenario (ratio=100), we visualise the uncertainty measurements (DE) of our method compared to vanilla EDL in Figure 5.6 (which can also be compared with Figure 5.1 as they share the same setup). Our method significantly improves the uncertainty quantification for minority classes such as class 1, 2 and 3, with larger uncertainty observed for incorrect predictions. In this case, the ROC-AUC for detecting incorrect predictions is 0.730. By setting a threshold of $DE = -60.0$, we can identify 75% of the incorrect predictions, greatly enhancing the robustness of the classifier by recognising what is unknown. Additionally, we display the uncertainty measurements for the in-distribution (ID) and out-of-distribution (OOD) testing sets in Figure 5.7. Similar to incorrect predictions, the uncertainty of the OOD set also increases as the distribution shifts from the left to the right. The ROC-AUIn this chapter, we delved into uncertainty-aware deep learning for multi-class health diagnostics. To enhance the effectiveness of the efficient uncertainty quantification method EDL, even in the presence of class-imbalanced physiological data, we introduced a class-balanced EDL approach with two



(a) Vanilla EDL. (b) Ours.

Figure 5.7: **Uncertainty distribution for CIFAR10**. Histograms of uncertainty measurements are presented for the application of OOD detection for a heavily imbalanced case.

innovative mechanisms. This results in fair and robust uncertainty estimates across all classes.C for detecting OOD in this example is 0.695 for our method and 0.625 for vanilla EDL.

In summary, the extended experimental results on the CIFAR10 benchmark validate the effectiveness of our proposed method in handling varying levels of imbalanced data, demonstrating improved performance and enhanced uncertainty quantification compared to vanilla EDL.

5.7 Discussion and conclusions

In this chapter, we delved into uncertainty-aware deep learning for multi-class health diagnostics. To enhance the effectiveness of the efficient uncertainty quantification method EDL in the presence of class-imbalanced physiological data, we introduced a class-balanced EDL approach with two innovative mechanisms for class imbalance.

Extensive evaluations using various real-world imbalanced physiological datasets substantiated the superiority of our method in achieving fair and robust uncertainty estimates across all classes. While we showed that our method achieved improved classification performance through uncertainty-aware learning, we also demonstrated the three key advantages of uncertainty quantification for health diagnostics in general: *(i)* Our method, designed to quantify model uncertainty, held significance in calibrating predictive probabilities for the in-distributional testing set. Even if the classification performance remained unchanged, the capability to reduce *ECE* enhanced the reliability of predicted confidence. *(ii)* Furthermore, such uncertainty information facilitated the identification of erroneous predictions, enabling their flagging for human review, especially by medical professionals. This ensured a proactive approach to preventing potential misdiagnoses for patients. This ability of our method was validated by the superior performance in AUC_m. *(iii)* Moreover, our method captured data uncertainty, effectively sig-

nalling instances of being out-of-training distribution. Therefore it substantially improved the reliability of deep learning models in real-world applications where data is often noisy and originates from diverse distributions. This was evidenced by the outstanding AUC_o achieved by our method.

Our evaluation may be subject to potential limitations arising from restricted data availability. Specifically, due to the small number of testing samples per class in the examined clinical tasks, we failed to assess class-level calibration results for the *ECE*. We intend to address this when more data becomes accessible. In the meantime, we suggest that future work consider using the class-agnostics *ECE* as the metric.

To conclude, our studies in Chapter 3 and Chapter 4 collectively contribute to a clearer understanding of a reliable automated system for health diagnostics. Such a system should be able to accurately detect changes in physiological status from sensory data. In addition to predicting individuals' health conditions, the system should provide accurate uncertainty estimations. If the prediction indicates high model uncertainty, the instance should be referred to doctors for timely diagnoses. In the event of high data uncertainty, the system should alert the individual to pay careful attention to the mobile device to minimise data noise and artifacts. Such capabilities can enhance the trustworthiness of automated health screening and promote better risk management without significantly increasing the burden on clinicians.

Chapter 6

FedLoss: Cross-device federated learning for distributed physiological data

Privacy is not an option, and it shouldn't be the price we accept for just getting on the Internet.

- Gary Kovacs

CEO at Accela Inc.

6.1 Introduction

As discussed in Chapter 1.2, privacy concerns surrounding health data pose a significant obstacle to the widespread sharing and centralised storage of large-scale physiological data for machine learning research. The recently emerged federated learning (FL) paradigm offers a promising solution for striking a balance between privacy protection and model development (Rieke et al., 2020; Li et al., 2020a; Nguyen et al., 2021a).

As detailed in Chapter 3.3.3, in contrast to traditional machine learning approaches, FL allows data to remain decentralised, residing where it was originally gathered. It facilitates global model training by aggregating model parameters instead of raw data. Specifically, starting with random initialisation, the global model is sent to clients (data holders), enabling them to optimise it using their local private data. After that, the server gathers the updated parameters from clients and aggregates them into a new global model. These two steps are repeated until the global model converges. This approach ensures that private individual data remains shielded from direct exposure, making federated learning exceptionally suitable for applications involving sensitive health-related data.

In this chapter, we aim to explore the possibility of developing models for health diagnostics using physiological data residing on mobile devices (Karimireddy et al., 2021). Under such a *cross-device* FL scenario, physiological data are distributed on individuals' mobile devices, such as smart wearables and smartphones, without sharing with any model developers. Each individual serves as a client, and they only exchange model parameters with the FL server to protect data privacy. However, this is a challenging task due to the following reasons: *i)* An individual's health status generally changes very slowly. Therefore, most personal devices will only present a single class, *i.e.*, the current health status of the device owner. It is infeasible to balance the data distribution on the device, and thus, learning from such data, the local model is likely to overfit and be biased. *ii)* Due to the generally low disease prevalence, the physiological data are also globally imbalanced, with a large proportion of healthy individuals. Without accessing the label distribution, global aggregation could introduce unwanted bias in the classification. Yet, failing to detect the disease may come at a heavy price in healthcare applications.

To address the local and global class imbalance problem, this chapter proposes an efficient federated training algorithm, *FedLoss*. The novelty of *FedLoss* lies in its adaptive model aggregation: only a small number of clients are required to participate in each round, and their models are aggregated according to adaptive weights proportional to the predictive loss on their local data. Such an adaptive aggregation strategy alleviates the impact of data imbalance and speeds up global model convergence. We validate the performance of *FedLoss* through the physiological audio-driven COVID-19 screening task (as introduced in Chapter 4.4). In this study, rather than using aggregated data, each participant is treated as an independent federated client, communicating only model parameters. Our experiments demonstrate that *FedLoss* achieves competitive performance compared to the centralised setting, indicating its effectiveness in handling data imbalance in the *cross-device* federated learning setting.

There are two main contributions made in this chapter.

- We propose a novel federated training algorithm to enable *cross-device* FL for mobile health diagnostics and tackle the challenge resulting from physiological data imbalance.

- We conduct extensive experiments in a real-world audio-driven COVID-19 detection task. Results demonstrate the superiority of our method over the start-of-the-art baselines.

The remainder of this chapter is organised as follows. We first review the related studies in Chapter 6.2. Then, we introduce our proposed method *FedLoss* in Chapter 6.3. Experimental setup and results are presented in Chapter 6.4 and 6.5, respectively. Finally, we conclude this chapter with a summary of our findings and a discussion of the limitations in Chapter 6.6.

6.2 Related work

Cross-device FL. Cross-device FL is a burgeoning field that addresses the challenges of decentralised data processing across multiple devices. In cross-device FL, a multitude of devices, often with varying capabilities and data profiles, collaborate to train machine learning models while keeping data localised, thus enhancing privacy and security (Karimireddy et al., 2021). Presently, there are billions of interconnected edge devices, including smartphones, tablets, and wearables, generating a continuous stream of data such as photos, videos, and audio (Lim et al., 2020). Such data presents numerous opportunities for meaningful research and applications. However, the conventional approach of aggregating this data in a central server is no longer sustainable, as the data can be sensitive to share and the communication cost associated with transferring such vast amounts of information can be prohibitive. Thanks to the advent of FL, developing models with data remaining at edge devices becomes feasible (Liu et al., 2021a).

Since one of the most compelling aspects of *cross-device* FL is its ability to leverage diverse data sources without requiring them to be aggregated in a central location, this approach is particularly beneficial in environments where data privacy is paramount, such as in healthcare and finance (ur Rehman et al., 2021). By allowing data to remain on users' devices, *cross-device* FL minimises the risk of data breaches and ensures compliance with stringent data protection regulations. However, *cross-device* FL is not without its challenges. The heterogeneity of devices can lead to issues such as imbalanced data contributions and computational disparities, which may affect the overall model performance and fairness (Karimireddy et al., 2021; Chen et al., 2023a). The data residing on personal devices can also be limited and non-representative, leading to biased local models. Additionally, managing communication efficiency in such a distributed setting is not trivial, as many devices are used for their functions and may not always be available for model training. (Yang et al., 2022a).

FedAvg, as introduced in Chapter 3.3.3, is the most commonly used algorithm for *cross-device* FL. It aggregates locally trained model parameters based on weights proportional to the fraction of data samples at clients compared to the total samples available in the system. However, *FedAvg* is susceptible to data distribution heterogeneity (Ma et al., 2022; Li et al., 2019; Zhao et al., 2018), a common issue in *cross-device* FL. To enhance real-world applications, several algorithms have been proposed. An extension of *FedAvg*, *FedProx*, adds a proximal term to the loss function used in training on local devices. This term helps address issues related to system heterogeneity, such as varying amounts of local data and different computational capabilities across devices. It generally yields better stability and convergence behaviour in heterogeneous environments, however, the choice of the proximal term's tuning parameter can significantly affect performance, requiring careful calibration (Li et al., 2020b). SCAFFOLD addresses the issue of client drift in federated learning, where updates from different devices diverge due to

having data that are not identically distributed (Karimireddy et al., 2020). It introduces control varieties that help correct the direction of the local updates towards the true gradient. SCAFFOLD improves learning efficiency and accuracy, particularly in non-IID data settings. Yet, it requires additional computation and communication overhead due to the control varieties. Some other efforts focus on client clustering (Lin et al., 2022a; Sattler et al., 2020), adapting the global model based on auxiliary data (Wang et al., 2021), and adaptive client training by monitoring loss from a global perspective (Zhang et al., 2021; Shen et al., 2021). However, these approaches are either inefficient with a large number of clients or require additional centralised data.

Cross-device FL for health diagnostics. In this chapter, we delve into the feasibility of developing a health diagnostics model using physiological data decentralised across mobile devices. Skewed label distribution across edge devices is a common occurrence in real-world applications, and some FL algorithms have been proposed as introduced above. However, for health diagnostics, the skewness can be even more severe (Rahman and Davis, 2013). As introduced at the beginning of this chapter, the health data residing on personal devices can be both locally and globally imbalanced. There are plenty of methods to address the class imbalance problem in a decentralised setting, as discussed in Chapter 3.3.1; however, they are not feasible for FL. Due to privacy constraints, handling class distribution cannot rely on explicitly identifying the minority class (Shen et al., 2021), rendering solutions explored in classical centralised settings invalid.

Currently, *cross-device* FL for health diagnostics remains largely under-explored in the literature. In a related study (Lin et al., 2022a) (*FedCluster*), a *cross-device* FL setting was considered for diagnosing arrhythmia from electrocardiograms. To enhance performance for the rare phenotype, *FedCluster* clusters clients based on a global shared dataset. Local models are then merged within clusters, and cluster models are aggregated into the global model. In contrast, we aim to address the imbalance problem without relying on any global data, and we make the first effort by proposing a novel solution.

There also have been a few studies on federated learning for COVID-19 detection, but *cross-silo* settings predominate with data distributed across multiple hospitals (Feki et al., 2021; Qayyum et al., 2022; Dou et al., 2021; Yang et al., 2021). For instance, Feki *et al.* proposed FL frameworks allowing multiple medical institutions to screen COVID-19 from Chest X-ray images without sharing patient data (Feki et al., 2021). Vaid *et al.* explored electronic medical records to improve mortality prediction across hospitals via FL (Vaid et al., 2021; Dayan et al., 2021). In these settings, the number of clients is small, and the size of local data is relatively large. To the best of our knowledge, we are the first to propose a *cross-device* federated learning framework for detecting COVID-19 from personal sounds and symptoms. This is more challenging than *cross-silo* FL due to the extreme data heterogeneity from thousands of clients.

6.3 Methodology

6.3.1 Problem formulation

In this section, we focus on the *cross-device* FL scenario and introduce our solution to address the physiological data imbalance challenge. For the sake of clarity, we first formulate the problem below. Following that, we introduce our proposed method, *FedLoss*.

Cross-device federated learning for health diagnostics. *Consider a FL system comprising K federated clients with each device, k, owning a private local dataset $\mathcal{D}^k = \{(x_k^{(1)}, y_k^{(1)}), ...\}$, where $x_k^{(i)}$ is a physiological data sample and $y_k^{(i)}$ denotes the health status, i.e., if the associated disease is identified in the sample, $y_k^{(i)} = 1$, otherwise $y_k^{(i)} = 2$. y_k is locally extremely imbalanced with most clients presenting a single class, and it is also globally imbalanced with $y_k = 2$ (healthy) being the majority class. In the end, we aim to train a federated model parameterised by θ that can predict y for any given x to achieve population health screening.*

6.3.2 FedLoss

> ***Highlight***: *In cross-device FL, model aggregation should take into account the local data distribution. The training loss serves as a valuable metric for determining the aggregation weights, enhancing the effectiveness of the aggregation process.*

As introduced in Chapter 3.3.3, *FedAvg* is the prominent FL algorithm (McMahan et al., 2017). It averages the model parameters, weighted by the fraction of local data sizes on the clients. However, this aggregation method fails to address the model bias present in physiological data, as it ignores imbalanced distribution. To address the limitation of *FedAvg*, we introduce an improved weighting mechanism to reduce the bias caused by imbalanced data distribution across clients (as detailed in the above hypothesis). Our proposed method, *FedLoss*, is summarised in Algorithm 1 and outlined as follows.

Sampling. Similar to *FedAvg*, the training of *FedLoss* proceeds iteratively. Before the iteration starts, the server randomly initialises a global model, termed by $\theta^{(0)}$ (r.f. line 2 in Algorithm 1). Starting from the initial model, clients will iteratively update the model parameters using local data (line 3-11 in Algorithm 1). In each round t, to make the learning scalable, we suppose M available clients are randomly selected to participate in the training process. The server will broadcast the parameters of the global model, $\theta^{(t-1)}$, to those selected clients.

Local training. As described in line 5-8 in Algorithm 1, each selected client, denoted as k, after receiving the global model, will optimises parameters for E epochs using its local data $\mathcal{D}_k$, and return the new models back the server. The details of local training are provided in line 12-22 in Algorithm 1. One significant difference between our method and *FedAvg* is that we propose

Algorithm 1: FedLoss Algorithm

Data: Global model update rate η, global training rounds T, local update rate λ, local training epochs E, the number of clients each round M.
Result: Global model $\theta^{(T)}$.

1 **Server executes:**
2 Initialise $\theta^{(0)}$
3 **for** *each round* t = *1,2,...,*T **do**
4 $\quad \mathcal{K}^{(t)} \leftarrow$ A random set of M clients
5 $\quad$ **for** *each client* $k \in \mathcal{K}^{(t)}$ *in parallel* **do**
6 $\quad\quad$ Send $\theta^{(t-1)}$ to client k
7 $\quad\quad$ $l_k^{(t)}, \theta_k^{(t)} \leftarrow k$-th client executes
8 $\quad$ **end**
9 $\quad$ Normalise weights: $w_1^{(t)}, ..., w_M^{(t)} = softmax(l_1^{(t)}, ..., l_M^{(t)})$
10 $\quad$ Model aggregation: $\theta^{(t)} \leftarrow \sum_{k=1}^{M} w_k^{(t)} \theta_k^{(t)}$
11 **end**
12 **Client** k **executes:**
13 Received a global model $\theta^{(t-1)}$
14 Initialise loss $l = 0$
15 **for** *sample* $j = 1, 2, ..., |\mathcal{D}_k|$ **do**
16 $\quad l \leftarrow l + \mathcal{L}(\theta^{(t-1)}; j)$ # Returning loss
17 **end**
18 Synchronise local model with the received parameters $\theta_k = \theta^{(t-1)}$
19 **for** *local epoch* $e = 1, 2, ..., E$ **do**
20 $\quad \theta_k \leftarrow \theta_k - \lambda \nabla_\theta \mathcal{L}(\theta_k; \mathcal{D}_k)$
21 **end**
22 Return l, θ_k

a new weight for model aggregation. We consider using the training loss which can reflect the dispensary between the model prediction and the ground truth. Thus, we calculate the predictive loss for the local data, as illustrated in line 15-17 in Algorithm 1. It is important to note that l is computed before the local training step, ensuring that it does not suffer from overfitting on a client with limited data. After deriving the predictive loss, then the model is optimised via back-propagation as introduced in Chapter 3.1.2 (line 18-21 in Algorithm 1). Local client k finally sends the predictive loss l and the updated model θ_k to the server.

Global aggregation. After local training, the server receives both the loss $l_k^{(t)}$ and the updated model parameters $\theta_k^{(t)}$. Since unhealthy clients are under-represented (globally minority class), intuitively they are more likely to yield relatively higher losses. To mitigate the bias, *FedLoss* will assign a higher weight to their model updates, rendering the data on such clients more predictable by the global model. Formally, the server normalises the received losses using a *Softmax* function to get the client-wise weights for aggregation (line 9-10 in Algorithm 1). The

adaptive aggregation in t-th round is denoted as,

$$\begin{aligned} w_1^{(t)}, ..., w_M^{(t)} &= Softmax(l_1^{(t)}, ..., l_M^{(t)}), \\ \theta^{(t)} &= \sum_{k=1}^{M} w_k^{(t)} \theta_k^{(t)}, \end{aligned} \tag{6.1}$$

where $w_k^{(t)}$ denotes the weight for the participating client k, and mathematically $w_k^{(t)} = \frac{exp(l_k^{(t)})}{1/M \sum_k exp(l_k^{(t)})}$.

After T rounds of iterations until the global model converges, we obtain the final model $\theta^{(T)}$, This model, abstracting the knowledge from all the clients' local data, now can be deployed to the population for health screening.

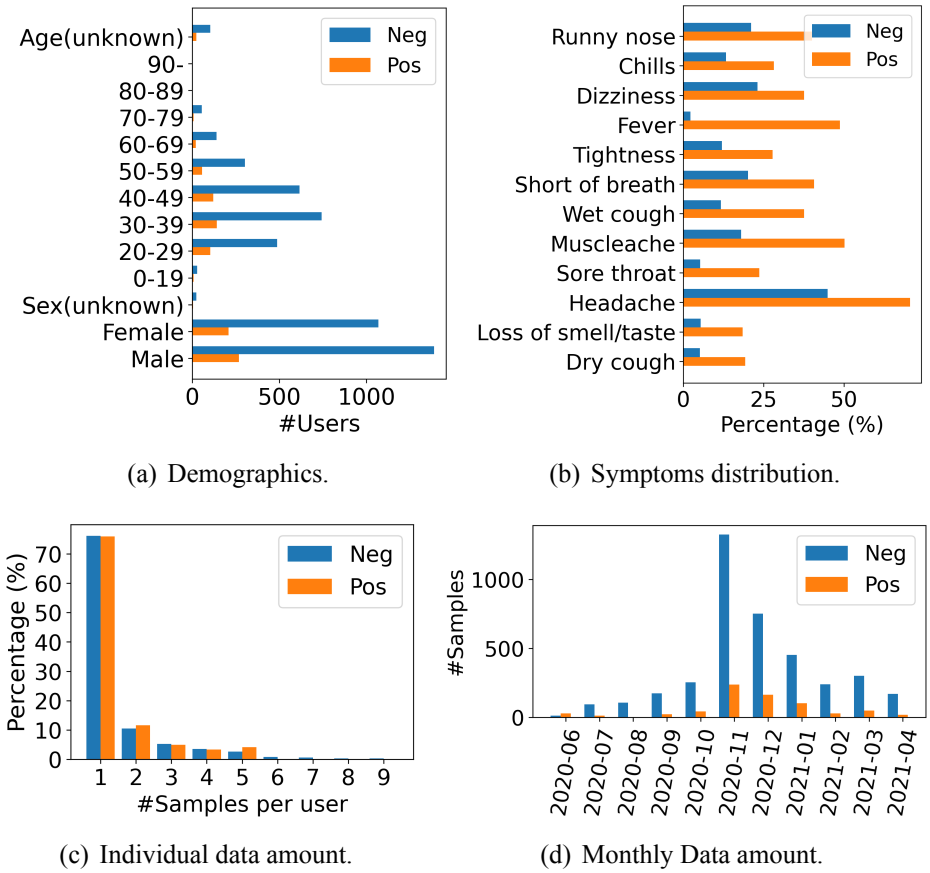
6.4 Experiments

Now we systematically evaluate *FedLoss* using the distributed physiological audio data. The following sections start with an illustration of our experimental setup and continue with a discussion of the results under two different settings.

6.4.1 Dataset and backbone model

For evaluation purposes, we again leverage the *COVID-19 Sounds* database as we used in Chapter 4 for the experiments. Similarly, with the setup in Chapter 4.4, we select English speakers and their samples with COVID-19 test results. But different from the study in Chapter 4.4, we include both symptomatic and asymptomatic positive participants, and we leverage the reported symptoms as input for the screening. Besides, for positive samples, we exclusively include those confirmed within the past 14 days. This decision is based on the consideration that a COVID-19 test conducted beyond 14 days may not accurately reflect the current infection at the time when the sounds were recorded.

As a result, there are 482 positive participants and 2,478 negative participants with a total of 4,612 samples. An overview of the statistics of the data is in Figure 6.1: (a) The data represents a representative demographic distribution in a population. (b) There are more negative than positive participants, with many asymptomatic positive participants while a great proportion of the negative participants report respiratory disease-related symptoms. (c) Individuals' data are sparse with over 70% of participants only recording one sample. (d) The data accumulation procession spanned one year. This dataset illustrates a typical decentralised physiological data distribution, characterised by global imbalance (more negative participants than positive participants) and local imbalance (the majority of participants recorded only one sample with a single COVID-19 status). Hence, this dataset is appropriate for evaluating *FedLoss*.



(a) Demographics.

(b) Symptoms distribution.

(c) Individual data amount.

(d) Monthly Data amount.

Figure 6.1: **Statistics of the data used for experiments**. All samples are from 482 COVID-19 positive participants and 2,478 negative participants.

Similar to the previously used model (refer to Figure 4.2 in Chapter 4.4), we utilise the pre-trained VGGish as the feature extractor for audio samples. Moreover, this model is a multi-modal one, taking both audio spectrograms and symptoms as inputs. The architecture of this multi-modal model is illustrated in Figure 6.2. Symptoms as illustrated in Figure 6.1(b) are represented by a 12-dimensional binary vector: the value for each dimension is set to 1 if the corresponding symptom is reported otherwise it equals 0. This symptom embedding is concatenated with the dense feature from VGGish outputs. The concatenated feature vector is then fed to a multi-layer fully connected network for classification. The final layer outputs a *Softmax* based binary class probabilities.

6.4.2 Federated learning setup

Out of 2,960 involved participants in the dataset we randomly hold out 20% participants for testing and use the rest 80% of the participant for federated training. We consider each participant as a federated client to examine *FedLoss*. We experiment with two training settings with different client training availability:

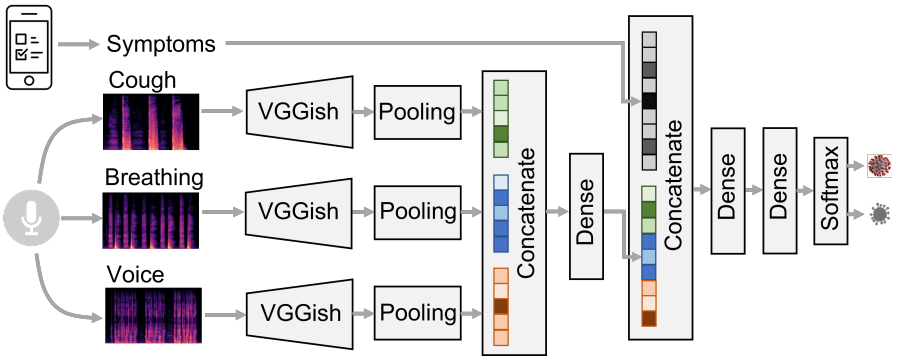


Figure 6.2: **COVID-19 screening backbone model**. The main architecture is adapted from Figure 4.2. The embeddings of spectrograms and the embedding for symptoms are concatenated to make more precise predictions for COVID-19 screening.

- **Randomly**: The recorded data is assumed to be kept on the client device during the whole training period. We run $T = 2000$ federated rounds and $M = 30$ clients are randomly selected at each round.
- **Chronologically**: The recorded data is assumed to be cleared monthly by the app user, which is practical. Regarding this, we design a multi-period training strategy: every month, only the clients with data recorded in this period can be selected and we run 100 rounds with each round sampling $M = 30$ clients for training (100 rounds can guarantee the convergence of the model on the incremental data).

Besides, for local training, the epoch is set to $E = 1$, and the fine-tuning learning rate is 0.008 for VGGish and 0.015 for the rest parameters.

6.4.3 Baselines and metrics

For this audio-based COVID-19 detection application, we first report the performance of centralised learning using the same training and testing sets. Centralised training yields the optimal performance that the FL methods can achieve, and thus serves as an upper bound. In addition to ***FedAvg***, we also compare with ***FedProx*** (Li et al., 2020b). *FedProx* handles non-identically distributed data across federated clients by regularising the local training loss at the clients so that the local models incur limited divergence from the global model. As discussed in Chapter 6.2, there are limited FL approaches for *cross-device* FL using health data. Thus, we mainly compared to *FedAvg* and *FedProx*.

For performance comparison, we report the following metrics: ***ROC-AUC***, ***Sensitivity***, and ***Specificity*** (as detailed in Chapter 3.4). Besides, we present the ***Youden's index*** (calculated as *Sensitivity* + *Specificity* - 1) to quantify the equilibrium between the rates of true positives and true negatives (Youden, 1950). The threshold for determining whether a sample is predicted to

Table 6.1: **Performance comparison under *randomly* shuffle setting**. 95% CIs are reported in brackets.

	ROC-AUC	Sensitivity	Specificity	Youden's index	SE@80%SP
Centralised	0.79 (0.74-0.84)	0.72 (0.66-0.76)	0.73 (0.68-0.76)	0.45 (0.40-0.50)	0.62 (0.54-0.69)
FedAvg	0.80 (0.75-0.85)	0.70 (0.65-0.74)	0.71 (0.65-0.75)	0.41 (0.36-0.44)	0.59 (0.45-0.73)
FedProx	0.78 (0.72-0.83)	0.70 (0.65-0.75)	0.70 (0.64-0.74)	0.40 (0.37-0.43)	0.48 (0.31-0.63)
FedLoss (Proposed)	0.79 (0.73-0.83)	0.72 (0.67-.76)	0.72 (0.68-0.7)	0.44 (0.40-0.47)	0.62 (0.50-0.70)

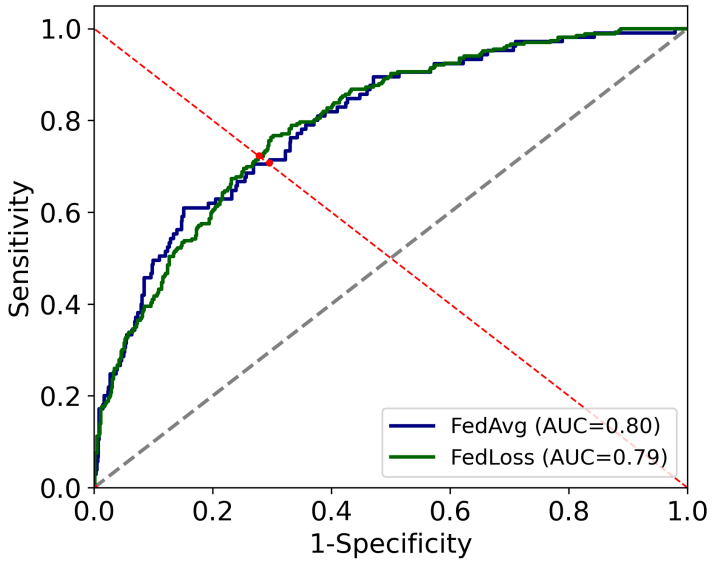


Figure 6.3: **ROC curves for FedAvg and FedLoss**. The threshold for determining whether a sample is predicted to be positive is identified on the ROC curve by balancing sensitivity and specificity, as shown by the red dots.

be positive is identified on the ROC curve of the testing set by balancing Sensitivity and Specificity. Additionally, we report ***SE@80%SP***, which is the sensitivity when using the decision threshold to ensure a specificity of 0.8. Furthermore, for both the baseline and our proposed methods, we report the 95% Confidence Interval (CI) for all metrics by using bootstrap (DiCiccio and Efron, 1996).

6.5 Results

Simulating the physiological audio data distributed locally on the original devices where they were collected, we now present the results for the two evaluated FL setups respectively.

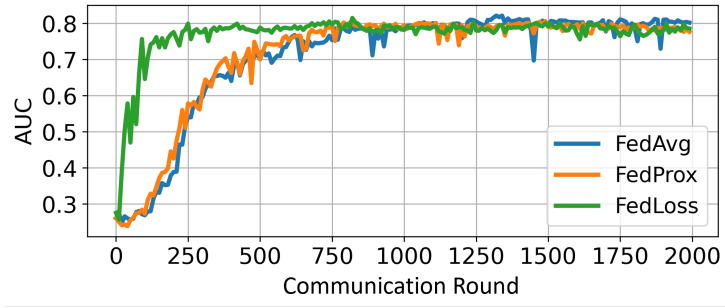


Figure 6.4: **Convergence analysis**. ROC-AUC of the testing set for every 10 rounds during training is displayed.

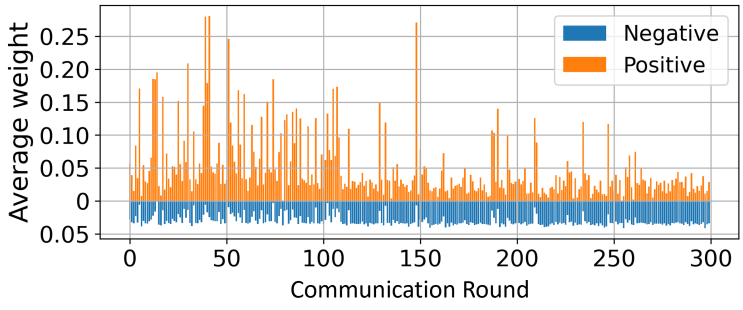


Figure 6.5: **Average weight for COVID-19 positive and negative clients per communication round**. Note that the negative clients do not have negative weights, but the weights are just shown in a negative direction for visualisation convenience.

6.5.1 Results under randomly shuffle setting

COVID-19 detection performance. The overall performance comparison is summarised in Table 6.1. All federated learning-based approaches achieve competitive ROC-AUC scores compared to centralised training. However, FedLoss achieves higher sensitivity and specificity than the federated baselines. When balancing sensitivity and specificity, FedLoss improves sensitivity by 2.9%. Specifically, the ROC curves of FedAvg and FedLoss are shown in Figure 6.3, where FedLoss yields a sensitivity of 0.72, compared to 0.7 for FedAvg. Furthermore, when optimising the sum of sensitivity and specificity, the Youden's index of FedLoss is 6.8% higher than that of the baselines. When specificity is fixed, the improvement increases to 5.1%. This validates the superiority of our weighted aggregation strategy in handling data imbalance.

Convergence comparison. System efficiency is another important metric for *cross-device* FL. To compare the convergence speed of *FedAvg*, *FedProx* and *FedLoss*, we show the testing ROC-AUC during the training process in Figure 6.4. It can be observed that the ROC-AUC of *FedLoss* converges significantly faster than the baselines: *FedLoss* needs about 250 rounds while *FedAvg* and *FedProx* requires about 1000 rounds. Therefore, *FedLoss* is $4\times$ more efficient than

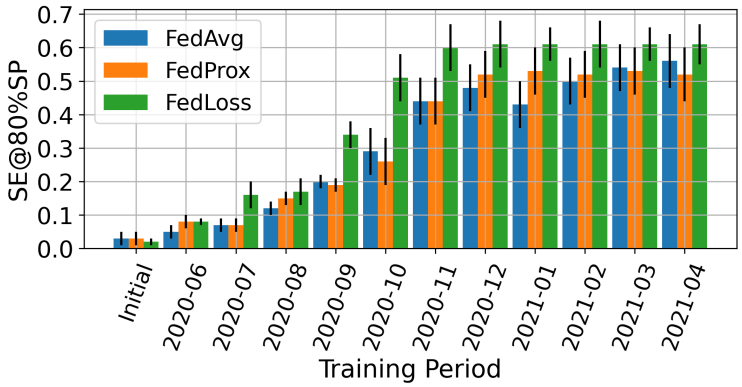


Figure 6.6: **Performance of the global model trained *chronologically***. Sensitivity when the specificity achieves 0.8 of the last round model in each month is displayed.

Table 6.2: **Overall performance under *chronologically* shuffle setting**. 95% CIs are reported in brackets.

	ROC-AUC	Sensitivity	Specificity	Youden's index	SE@80%SP
FedAvg	0.79 (0.73-0.82)	0.70 (0.65-0.73)	0.70 (0.65-0.75)	0.40 (0.35-0.44)	0.56 (0.49-0.63)
FedProx	0.78 (0.75-0.81)	0.69 (0.64-0.73)	0.69 (0.64-0.73)	0.39 (0.34-0.43)	0.53 (0.44-0.60)
FedLoss (Proposed)	0.79 (0.74-0.84)	0.71 (0.66-0.76)	0.72 (0.67-0.76)	0.43 (0.39-0.47)	0.61 (0.55-0.64)

the baselines. Note that fewer communication rounds to converge save both computation and communication costs at the resource constraint edge clients (Wu et al., 2022).

Analysis of weights. We conducted additional analysis on the adaptive weights during the training process. Given that *FedLoss* exhibits superior sensitivity, we specifically examined how the weights changed for COVID-19 positive and negative clients for comparison. Figure 6.5 illustrates the average weight for positive and negative clients in each round.

It is noteworthy that in the first 100 rounds, the weight of positive clients is approximately 4 to 6 times that of negative clients. This implies that the system can effectively identify the under-represented class, as these clients are more challenging to predict. In the subsequent rounds, the weights for positive and negative clients gradually balance out, indicating that the global model has already assimilated COVID-19 features to a significant extent.

However, it is essential to note that this trend does not necessarily imply that positive participants consistently have higher weights than negative participants. As illustrated in Figure 6.5, it is evident that many positive participants have weights as small as those of negative participants. This observation indicates that the weight distribution does not disclose class information.

6.5.2 Results under chronologically training setting

The second setting aims to evaluate the performance of long-term FL with limited client participation in batches. As illustrated by SE@80%SP in different periods in Figure 6.6, all methods are inaccurate and unusable at the early stage with an index lower than 50%. The poor performance is mainly attributed to the limited number of clients (*i.e.,* the limited data), which leads to poor generalisation. Gradually, with more training rounds, from November 2020 *FedLoss* starts to show a convergence trend with SE@80%SP reaching 60%. Finally, our model achieves a ROC-AUC of 0.79, a sensitivity of 0.71 and a specificity of 0.72, as summarised in Table 6.2. On the contrary, SE@80%SP of *FedAvg* and *FedProx* have a slower convergence rate, converging two months later than *FedLoss*. We also note that from November 2020, all three approaches present a remarkable performance gain, which is mainly because the quantity of data reaches a peak in that month (refer to Figure 6.1(d)). Overall, our final Youden's index (0.43) surpasses that of *FedAvg* (0.40) and *FedProx* (0.39), and our sensitivity (0.71) is quite competitive compared with the centralised model (0.72). The above comparison further verifies that our proposed *FedLoss* can achieve a more generalisable global model with fewer clients involved.

Overall, experimental results demonstrate the effectiveness of *FedLoss* with different levels of training data availability. *FedLoss* shows a higher sensitivity in handling the data imbalance issue in the *cross-device* FL scenario, a challenge that *FedAvg* and *FedProx* fail to address. Our work paves the way for transitioning from traditional crowd-sourcing of data to crowd-sourcing of model parameters on a large scale, enabling privacy-preserving mobile health research.

6.6 Discussion and conclusions

This chapter explored the potential of utilising decentralised physiological data for deep learning model development through federated learning. To tackle challenges arising from locally and globally imbalanced physiological data across mobile devices, we introduced a novel *cross-device* federated learning method called *FedLoss*, enhancing the accuracy and fairness of federated learning in mobile health applications. Experimental results on a real-world COVID-19 screening task demonstrated the superiority of our approach in terms of both effectiveness and efficiency.

One limitation of our work is that *FedLoss* specifically targets a binary diagnostic application. Its general applicability for multi-class health diagnostics remains a subject for future investigation. While we evaluated our method using the physiological audio database, *FedLoss* is agnostic to data modality and model architecture. We envision its potential deployment across various mobile health applications, such as arrhythmia prediction based on heart sounds and

monitoring sleep quality through smartwatches. These areas represent opportunities for future exploration. Beyond the immediate impact of our work, we see our contribution fostering a shift from traditional data crowd-sourcing to large-scale crowd-sourcing of models in the domain of privacy-preserving mobile health research.

Another assumption in this chapter is that mobile devices can train a deep-learning model using their local data. Our experiments involve simulating communication between clients and the FL server; however, real-world mobile devices may face constraints in computation, making it impractical to train the model. Recent research on device learning has been addressing challenges related to energy consumption, memory constraints, and latency issues (Dhar et al., 2021; Lin et al., 2022b; Cai et al., 2020). We are optimistic that *cross-device* federated learning will soon transition from theory to reality.

Chapter 7

FLea: Cross-silo federated learning for distributed physiological data

Alone we can do so little; together we can do so much.

- Helen Keller

- An author, activist, and lecturer

7.1 Introduction

Physiological data, such as ECG and EEG signals, are extensively collected and annotated in health institutions like hospitals. However, relying solely on data from a single resource may prove insufficient for developing high-performance health screening models due to limitations in data scale and the ability to encompass a comprehensive range of health conditions, especially for rare diseases. The conventional approach involves aggregating anonymous data from multiple resources to a central server, with restricted access granted only to a few stakeholders, as introduced in Chapter 1.2. Unfortunately, this process could hinder both data collection and model development, as indicated by previous studies (Crow et al., 2006; Kreuter et al., 2020). Furthermore, the risk of malicious attacks during data transmission or storage on the server remains a significant concern (Li and Liu, 2021).

As elucidated in Chapter 3.3.3, the advent of federated learning (FL) presents a promising solution, enabling multiple institutions to collaboratively train a model without sharing their data—referred to as *cross-silo* FL (Huang et al., 2022a). However, the inherent data heterogeneity across these multiple resources poses a significant challenge to the performance of FL. Specifi-

cally, discrepancies in demographics, disease prevalence, and other data collection-specific factors result in heterogeneous data distributions across these institutions. This data heterogeneity also implies that local datasets are insufficient for accurately representing the health data distribution at the population level. As introduced in Chapter 3.3.3, when trained solely on such local data, the local model may deviate from the optimal model, potentially diminishing the performance of the aggregated global model.

To address the issue arising from heterogeneous local data distributions, in this chapter, we introduce a novel method called *FLea*. In addition to sharing model parameters (*i.e.*, the typical FL approach), *FLea* encourages clients to exchange privacy-protected features alongside model parameters to assist in local training. These features are derived from activations in an intermediate layer of the model, which are obfuscated before sharing with other clients to protect sensitive information in the data. We propose a new approach to combine local and shared features as augmentations for local model training. This can alleviate model drift caused by local data discrepancies and enhance the performance of the global model.

In our experiments, we first verify the superiority of our method using multi-centre ECG data: *FLea* achieves competitive accuracy compared to the model trained with centralised data. We further evaluate *FLea* via distributing a general machine learning benchmark dataset into multiple data silos with various levels of heterogeneity. Results show that *FLea* outperforms state-of-the-art FL counterparts, which share only model parameters, by up to 17.2%, and FL methods that share data augmentations by up to 6.2%, while also mitigating the privacy vulnerabilities in shared data augmentations.

Overall, this chapter makes the following contributions:

- We propose a novel *cross-silo* FL approach, *FLea*, to address the client data heterogeneity problem. To the best of our knowledge, *FLea* is the first FL method that leverages globally shared and privacy-preserving features as data augmentations.
- We evaluate the performance of *FLea* using real-world multi-centre ECG data, showcasing its effectiveness in developing a deep learning model for health diagnostics without aggregating physiological data from multiple resources.
- We also conduct extensive experiments on machine learning benchmark data. The results not only suggest the superior performance of *FLea* compared to state-of-the-art FL baselines but also highlight its privacy-serving advantages over existing data-sharing-based FL counterparts.

The remainder of this chapter is organised as follows. We first review related studies in Chapter 7.2. Then, we introduce our proposed method, *FLea*, in Chapter 7.3. The experimental setup is presented in Chapter 7.4. We discuss the results on the multi-centre ECG data in Chapter 7.5

and the machine learning benchmark in Chapter 7.6, respectively. Finally, we conclude this chapter with a summary of our findings and a discussion of the limitations in Chapter 7.7.

7.2 Related work

We have introduced the most widely adapted FL method *FedAvg* in Chapter 3.3.3. It aggregates locally trained model parameters based on weights proportional to the fraction of data samples at clients compared to the total samples available in the system. However, *FedAvg* is susceptible to data distribution heterogeneity (Ma et al., 2022; Li et al., 2019; Zhao et al., 2018). To promote the real-world application of *cross-silo* FL, many more advanced algorithms have been proposed in the broader field of machine learning, which can be mainly categorised into three types: *data-based*, *loss-based*, and *aggregation-based* methods as follows.

Data-based methods. These methods rely on the assumption and use of a global shared dataset, which can be either collected or generated, to align local models. *Zhao* et al. proposed a method called *FedData*. This method demonstrates that sharing a small proportion of local data globally, alongside the model parameters, can effectively enhance *FedAvg* (Zhao et al., 2018). Despite the desirable performance gains brought by *FedData*, collecting private data would compromise the privacy-preservation benefits of FL. Other global proxies less privacy-sensitive than raw data have been proposed to augment the local data. *FedMix* (Yoon et al., 2020) and *FedBR* (Guo et al., 2023) average data over mini-batches and share this aggregated data globally. Data generation methods are also explored in FL, such as learning a data generator at the server (Liu et al., 2022) or locally at the client (Jeong et al., 2018). However, the quality of such generated data is typically insufficient to enhance the final performance. In response to this challenge, a post-hoc approach called *CCVR* was proposed, which fine-tunes the classification layer of the global model using global deep features (Luo et al., 2021). Nonetheless, we have two concerns: (i) the accumulation of local model drift over time, resulting in only marginal performance gains from post-hoc calibration, and (ii) the limitation of potential performance gain by only tuning the classification layer. Despite these limitations, this method served as inspiration for the design of *FLea* as a feature-sharing method for cross-soil federated learning. *FLea* introduces novel strategies to address the limitations of *CCVR*.

Loss-based methods. These methods regularise local training to force that the locally trained model remains close to the globally shared model. For example, *FedProx* uses a penalty quantified by the difference between the global and local model parameters along with the local data learning loss to prevent local model drift (Li et al., 2020b). Further, Yu *et al.* proposed to learn the representations merely from private data while keeping the classification layer frozen (Yu et al., 2020) *SCAFFOLD* leverages the similarity between clients to reduce the variance of model updates (Karimireddy et al., 2020). *FedRS* (Li and Zhan, 2021) and *FedLC* (Zhang et al., 2022)

calibrate the logits for the local absent classes during local training and thus prevent local models from drifting away. Recently conducted studies have shed light on the impact of heterogeneous data on local forgetting (Liu et al., 2022; Shoham et al., 2019; Lee et al., 2022). Along this way, *FedNTD* (Lee et al., 2022) aims to leverage global knowledge from the global model to prevent local forgetting.

Aggregation-based methods. Aiming to directly enhance *FedAvg*, aggregation-based methods are proposed to effectively aggregate local models, avoiding a sole reliance on fixed weights. Yeganeh *et al*. introduced inverse distance aggregation to improve *FedAvg* (Yeganeh et al., 2020). The essence of this method lies in the computation of weights, based on the inverse distance of each client's parameters to the average model of all clients. This approach allows the server to reject or assign lower weights to models that may potentially poison the system, such as out-of-distribution models. While this type of aggregation enhances the robustness of FL, it is limited to addressing the inner data distribution heterogeneity. *FedLoss*, introduced in Chapter 6, is also an aggregation-based FL approach. However, it is specifically applicable to binary classification tasks, whereas *cross-silo* FL typically deals with multi-class data.

In this chapter, our objective is to devise a method that integrates both data-based and loss-based approaches to address the challenge of data heterogeneity in *cross-silo* federated learning. In contrast to the data-based approaches mentioned earlier, our method seeks to minimise privacy exposure linked to shared data augmentations. Additionally, we introduce a novel training loss to efficiently leverage globally shared data augmentations for local training.

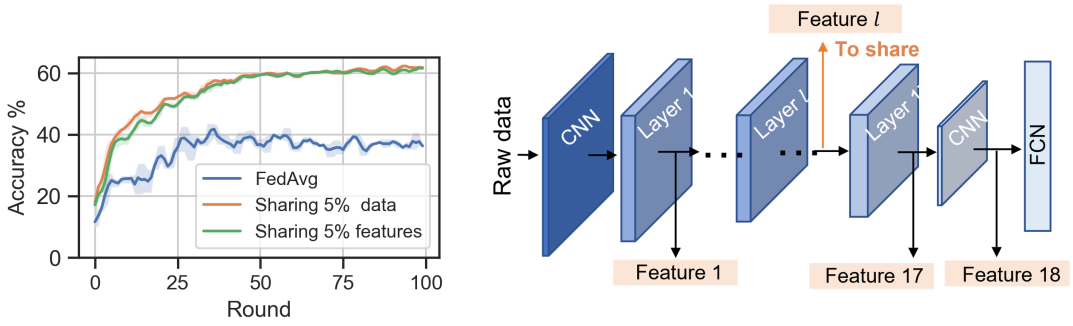
7.3 Methodology

7.3.1 Problem formulation

Our studied *cross-silo* FL problem is formulated as follows:

Cross-silo federated learning for health diagnostics. *K health institutes having physiological datasets aspire to collaboratively develop a health diagnostic model without sharing their data. However, each local dataset $\mathcal{D}_k$ contains varying amounts of different subsets of the globally available classes $\mathcal{C}$ (label distribution heterogeneity) and the size of $\mathcal{D}_k$ varies across the institutes (data quantity heterogeneity). Under such data distribution, the goal is to develop a global model θ that can accurately diagnose all health conditions that presents in the K datasets.*

To tackle the above-defined problem, we introduce a novel method, *FLea*. Before delving into our proposed method to address the challenges in *cross-silo* FL, we present some insights that inspired our work as follows.



(a) Overall accuracy.

(b) Feature extraction.

Figure 7.1: **Results for CIFAR10 classification under label distribution heterogeneity.** In (a), the performance of *FedAvg* and the performance with globally sharing 5% of the data and 5% of the features are compared. (b) illustrates a model and the features extracted from its middle layers. When Feature 1 is shared globally, it achieves similar performance with sharing the same amount of raw data, as can be observed in (a).

7.3.2 Motivation

As discussed earlier, the presence of local data heterogeneity poses significant challenges to the effectiveness of FL. To delve into this problem, the heterogeneous local distribution hinders high-performing FL for the following reasons: 1) Local models exhibit bias due to the absence of certain classes in the local dataset $\mathcal{D}_k$, rendering them ineffective for the global distribution; 2) Local models differ from each other, and simply aggregating them cannot retain useful knowledge. Addressing these problems is crucial for enhancing FL in the presence of label distribution heterogeneity in decentralised physiological data (Zhao et al., 2018; Zhu et al., 2021; Guo et al., 2023).

Existing works have demonstrated that globally sharing a small portion of the data can be highly effective in addressing label distribution heterogeneity. Training local models with a global proxy containing all classes helps mitigate local biases and align local models simultaneously. We use a CIFAR10 classification example to illustrate this. In this example, the CIFAR10 data (introduced in Chapter 2.2.3) is distributed among 10 clients, each having three out of the overall ten classes. As shown in Figure 7.1(a), in the presence of data heterogeneity, by globally sharing just 5% of the data, the FL model's accuracy improved by over 20% compared to *FedAvg* which does not share any data. In this case, *FedAvg* achieves an accuracy on the testing set with ten classes of about 40%, while globally sharing 5% of local data can boost the performance to 60%.

Despite the desirable performance gains offered by this data-sharing method in healthcare applications, sharing private physiological data with health condition labels to others would compromise the privacy-preservation target. Therefore, a natural question arises: *Can we share some information that can yield similar performance gains as sharing a subset of raw data, but with*

reduced privacy risks?

To this end, in this section, we propose to exchange deep features to aid local training. Those features correspond to the outputs of an intermediate layer in the deep neural network, which receives local data as input, as depicted in Figure 7.1(b). Deep models are typically composed of multiple layers of interconnected artificial neurons that learn to extract increasingly abstract features from input data. Thus, features not only are meaningful for classification but also provide an opportunity to protect the privacy associated with raw data (Vepakomma et al., 2020). Given this, we introduce a novel privacy-preserving feature sharing and augmentation method, namely *FLea*, for cross-silo federated learning.

7.3.3 FLea

***Highlight**: In cross-silo FL, a global proxy has the potential to prevent the local model from drifting when dealing with heterogeneous local data distributions. We hypothesise that features extracted from the intermediate layers of the model can serve as a robust proxy to enhance learning while maintaining privacy protection.*

To address local data heterogeneity, the main idea behind *FLea* is achieved by exchanging features among clients along with the model parameters. As illustrated in Figure 7.2, there is a global feature buffer that aggregates feature-label pairs from multiple clients, serving as a global proxy to assist local training. To make full utilisation of the feature buffer, *FLea* introduces a novel feature augmentation approach to combine local and global features. Additionally, a knowledge distillation strategy is applied to the combined features to further prevent local model drift. To protect data privacy, features are shared by clients after applying a certain level of "obfuscation": we reduce the correlation between the features and the data while maintaining their classification characteristics through a customised loss function.

Concretely, *FLea* works in a iterative manner, similarly to *FedAvg* (refer to Chapter 3.3.3) and *FedLoss* (the method we proposed in Chapter 6). In *FLea*, the server will maintain a global model and a feature buffer which contains feature-target pairs from multiple clients. In the beginning, the global model is randomly initialised and the buffer is empty. Then as illustrated in Figure 7.2, *FLea* works iteratively to update the global model and the buffer. Each round of *FLea* starts with synchronising the global model parameters $\theta^{(t)}$ and feature buffer $\mathcal{F}^{(t)}$ to the K clients. Once local training using $\mathcal{D}_k$ and $\mathcal{F}^{(t)}$ finishes (in the first round, only local data $\mathcal{D}_k$ is used for training since the feature buffer is empty), those clients send the updated model parameters θ_k to the server, to be aggregated into a new global model parameterised by $\theta^{(t+1)}$. *FLea* uses the same aggregation strategy as *FedAvg* (Eq. (3.12)). Followed by that, *FLea* needs another step to update the global feature buffer to $\mathcal{F}^{(t+1)}$. A detailed training procedure can be found in Algorithm 2. We elaborate on the main components of the procedure as follows.

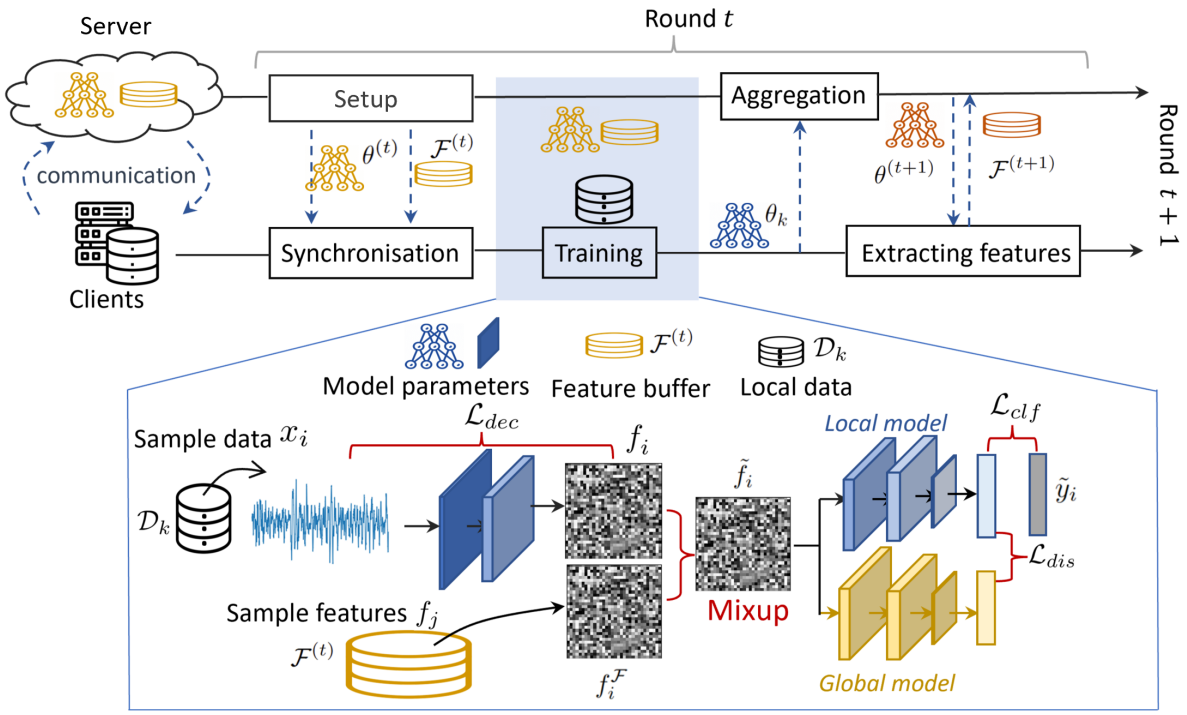


Figure 7.2: **An overview of FLea**. The training process for t-th communication round is shown.

Feature buffer: Let us consider the global model parameters $\theta^{(t)}$ to be divided into two parts at layer l: $\theta^{(t)}[:l]$ and $\theta^{(t)}[l:]$. For client k, the feature vector extracted from a data point $x_i \in \mathcal{D}_k$ is $\theta^{(t)}[:l](x_i) = f_i^{\mathcal{F}}$. The feature buffer from this client is the set of pairs including target labels and feature vectors $(f_i^{\mathcal{F}}, y_i^{\mathcal{F}})$. Each client randomly selects a α fraction of its local data to create its feature buffer to share with others. The server gathers those local feature buffers and merges them into the global one $\mathcal{F}^{(t)}$. Note that a client only extracts and contributes to the global feature buffer at the round when it participates in training and the global buffer resets at every round.

Client k's local training: Now let us look at the local training for client k in round t. As shown in Figure 7.2, k receives the global model $\theta_k = \theta^{(t)}$ and the feature buffer $\mathcal{F}^{(t)}$. The local data $\mathcal{D}_k$ and the feature buffer $\mathcal{F}^{(t)}$ are divided into equal-sized batches for model optimisation, termed by $\mathcal{B} = \{(x_i, y_i) \in \mathcal{D}_k\}$ and $\mathcal{B}^f = \{(f_i^{\mathcal{F}}, y_i^{\mathcal{F}}) \in \mathcal{F}^{(t)}\}$, respectively ($|\mathcal{B}| = |\mathcal{B}^f|$). The traditional method will feed $\mathcal{B}$ into the model directly to optimise the model but we propose to augment the input in the feature space. We feed $\mathcal{B}$ into the model, extracting the intermediate output for each data point: $f_i, \forall x_i \in \mathcal{B}$, and generate the augmentation as,

$$\begin{aligned} \tilde{f}_i &= \lambda_i f_i + (1-\lambda_i) f_i^{\mathcal{F}}, \\ \tilde{y}_i &= \lambda_i y_i + (1-\lambda_i) y_i^{\mathcal{F}}, \end{aligned} \tag{7.1}$$

where (f_i, y_i) and $(f_i^{\mathcal{F}}, y_i^{\mathcal{F}})$ are two feature-target pairs from $\mathcal{B}$ and $\mathcal{B}^f$, respectively. Inspired

Algorithm 2: Federated Learning with Feature Sharing (FLea)

Input : Total rounds T, local learning rate η, local training epochs E, clients wit local data $\mathcal{D}_k$, a given layer l, parameter a for Beta distribution.
Output: Global model $\theta^{(T)}$.

Initialise $\theta^{(0)}$ **for the global model**
for *each round* t = *1,2,...,*T **do**
  Server broadcasts the feature buffer $\{\mathcal{F}^{(t)}, .., \mathcal{F}^{(t-\tau)}\}$ to clients // Skip if $t = 1$.
  for *each client* k *in parallel* **do**
    Server broadcasts $\theta_k \leftarrow \theta^{(t-1)}$
    for *local step* $e = 1, 2, .., E$ **do**
      for *local batch* $b = 1, 2, ...$ **do**
        sample $\lambda \sim Beta(a, a)$
        $\theta_k \leftarrow \theta_k - \eta \nabla \mathcal{L}(\theta_k)$ // if $t = 1$, only use local data for training. Otherwise, use one batch of local data $\mathcal{D}_k$ and one batch of global feature $\mathcal{F}^{(t)}$ according to Eq. (7.5).
      end
    end
    Client k sends θ_k to server
  end
  Server aggregates θ_k to a new global model $\theta^{(t)}$ refer to Eq. (3.12)
  for *each client* k *in parallel* **do**
    Client k receives model $\theta^{(t)}$
    Client k extracts (without gradients) and sends $\mathcal{F}_k^{(t)}$ to server
  end
end

by the data augmentation method in the centralised setting (Zhang et al., 2018), we sample the weight λ_i for each data point from a symmetrical Beta distribution (Gupta and Nadarajah, 2004): $\lambda_i \sim Beta(a, a)$. $\lambda_i \in [0, 1]$ controls the strength of interpolation between the local and global feature pairs: A smaller λ_i makes the generated sample closer to the local feature while a larger one pushes that to the global feature.

One might question the advantage of employing such augmentation in comparison to directly combining local and global feature-target pairs for local model training. We posit that the augmentation outlined in Eq. (7.1) can contribute to performance enhancement in the following ways: (1) It transforms hard-label optimisation into soft-label optimisation, where y_i and $y_i^{\mathcal{F}}$ represent hard-labels while $\tilde{y}_i$ signifies soft-labels. This transformation has the potential to mitigate local overfitting, enhancing the generalisability of local models, and thereby facilitating a better global model. (2) The utilisation of random sampling and mixup augmentation introduces diversity by generating additional training data points. To elaborate, each training batch becomes unique as the weighting rate λ is randomly sampled.

Following the augmentation, the training loss for each batch is designed to contain two parts:

one for classification ($\mathcal{L}_{clf}$) and one for knowledge distillation from the global model ($\mathcal{L}_{dis}$). The classification loss $\mathcal{L}_{clf}$ is formulated as,

$$\mathcal{L}_{clf}(\mathcal{B}, \mathcal{B}^f) = \frac{1}{|\mathcal{B}|} \sum_i \sum_c -\tilde{y}_i[c] \log p_i^l[c], \tag{7.2}$$

where for $\tilde{f}_i$, the *logit* is $z_i^l = \Gamma_{\theta_{k,l:}}(\tilde{f}_i)$ and the probability for class c is $p_i^l[c] = \frac{exp(z_i^l[c])}{\sum_c exp(z_i^l[c])}$. The distillation loss (Hinton et al., 2015) is derived by the KL-divergence between the global probabilities and local probabilities as,

$$\mathcal{L}_{dis}(\mathcal{B}, \mathcal{B}^f) = \frac{1}{|\mathcal{B}|} \sum_i \sum_c -p_i^l[c] \log \frac{p_i^g[c]}{p_i^l[c]}, \tag{7.3}$$

where for $\tilde{f}_i$ the global *logit* is $z_i^g = \Gamma_{\theta_{l:}^{(t)}}(\tilde{f}_i)$ and the global probability is $p_i^g[c] = \frac{exp(z_i^g[c])}{\sum_c exp(z_i^g[c])}$. Meanwhile, we aim to obfuscate the features to protect data privacy before they are shared with the clients. As such, we learn the l layers while reducing the correlation between the features and the source data. This is achieved by minimising the loss (Vepakomma et al., 2020) below,

$$\mathcal{L}_{dec}(\mathcal{B}) = \frac{Tr(X^T F F^T X)}{\sqrt{Tr(X^T X)^2}\sqrt{Tr(F^T F)^2}}, \tag{7.4}$$

where $X \in \mathbb{R}^{|\mathcal{B}| \times d}$ and $F \in \mathbb{R}^{|\mathcal{B}| \times d^f}$ are the data and feature matrix. Note that each $X_i \in \mathbb{R}^d$ and $F_i \in \mathbb{R}^{d^f}$ are the flattening vector for data x_i and feature f_j^l. The numerator of $\mathcal{L}_{dec}$ measures the covariance between the data and the features, while its denominator measures the averaged pairwise distance within the data batch and feature batch, respectively. When updating the local model, the features change correspondingly. It is desired that the distance covariance decreases faster than the feature inner distance for each batch. Since when reducing the correlation, we hope the features can maintain classification ability, and thus we optimise all the loss functions jointly, as follows,

$$\mathcal{L} = \mathcal{L}_{clf}(\mathcal{B}, \mathcal{B}^f) + \lambda_1 \mathcal{L}_{dis}(\mathcal{B}, \mathcal{B}^f) + \lambda_2 \mathcal{L}_{dec}(\mathcal{B}), \tag{7.5}$$

where λ_1 and λ_2 are the weights to trade-off classification and privacy preserving. The local update is then achieved by $\theta_k \leftarrow \theta_k - \eta \frac{\partial \mathcal{L}}{\partial \theta_k}$, where η controls the learning rate.

Feature buffer updating: After the global model aggregation and broadcasting, client k extracts the features from the new model parameterised by $\theta^{(t+1)}$ from layer l to formulate the feature set. Those sets will be sent to the server to replace the old ones, updating the feature buffer to $\mathcal{F}(t+1)$. The iterations continue until the global model converges.

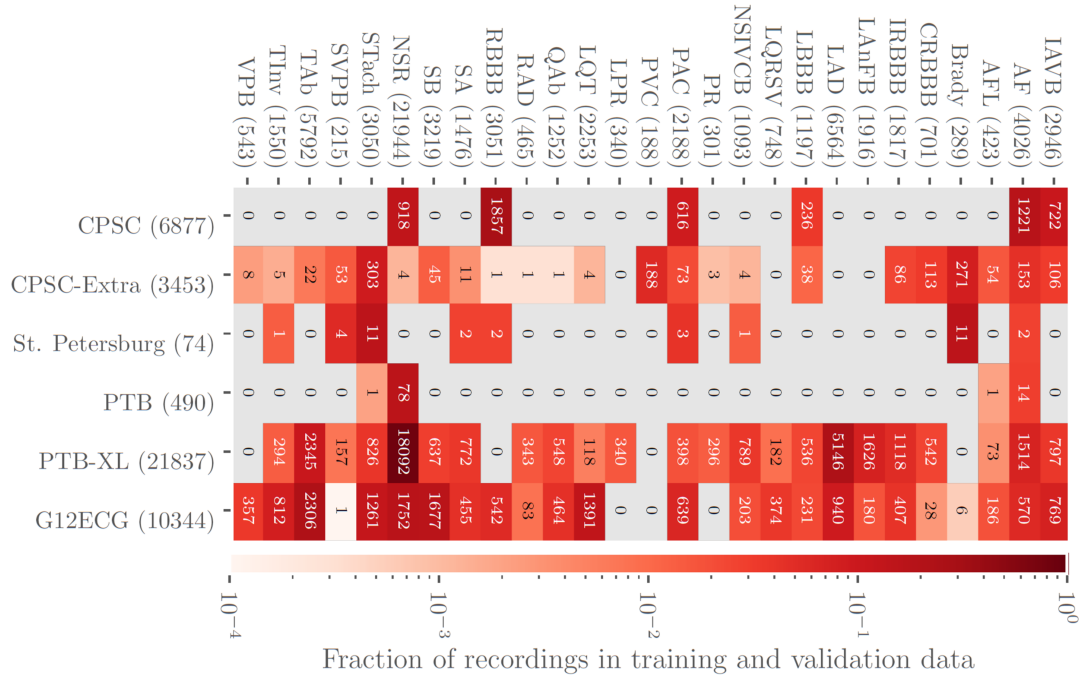


Figure 7.3: **Numbers of recordings with each scored diagnosis across data silos**. Colours indicate the fraction of recordings with each scored diagnosis in each data silo, *i.e.*, the total number of each scored diagnosis in a silo normalised by the number of recordings in each data set. Parentheses indicate the total number of records with a given label (Alday et al., 2020).

7.4 Experimental setup

To evaluate our method, extensive experiments are conducted on a real-world ECG database and a machine learning benchmark. The setups as introduced below.

7.4.1 Setup for ECG data

Dataset. We leverage the multi-centre ECG database *CIC2020* for the federated learning experiments. Detailed information about the data can be found in Chapter 2.2.2. The data was collected from 6 different institutes and were annotated with a total of 27 cardiovascular abnormalities, the distribution of which is provided in Figure 7.3. Notably, the data exhibit significant heterogeneity in both the amount of data, ranging from 74 to 21,837 recordings, and the number of classes, varying from 5 to 23 for each silo. Therefore, this database is well-suited for evaluating our method.

ECG classification model. We employ a ResNet-based architecture for the classification of ECG recordings (He et al., 2016). The effectiveness of this architecture has been previously demonstrated via a centralised 12-lead ECG data (Yang et al., 2020). ResNet was originally proposed for feature extraction from image data. For ECG classification, Yang *et al.* replaced the previous 2-dimensional convolutional kernel with a 1-dimensional convolutional kernel con-

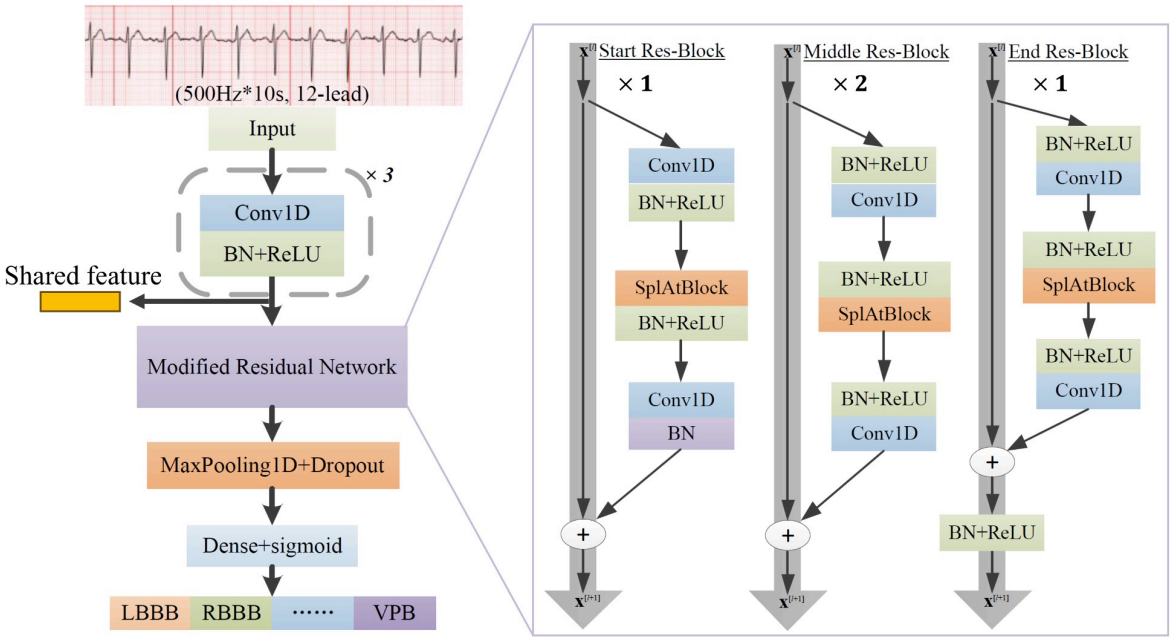


Figure 7.4: **ECG classification model based on the modified residual convolutional network**. The proposed model is mainly composed of multiple basic blocks and four modified residual convolutional network stages, as shown on the right side. The split attention block (SplAtBlock) in Res-Block divides the feature into several feature-map groups and the combined representation of each cardinal group can be obtained by fusing via an element-wise summation across multiple splits.

sidering the time series characters of ECG. The overall model is illustrated in Figure 7.4. Yang *et al.* introduced modifications to the network architecture for residual learning, creating better paths for information to propagate through the network layers (Yang et al., 2020). Finally, the learned features are passed through fully connected layers with *Sigmoid* activation to achieve multi-class classification.

The ECG data from six different resources were collected at varying sampling rates and presented in different lengths. To ensure a uniform input for the model, we perform data preprocessing. First, the data from four different sources were sampled at frequencies of 257 Hz, 500 Hz, and 1000 Hz. To maintain data consistency as much as possible, we re-sample the ECG records with frequencies of 257 Hz and 1000 Hz to 500 Hz. Second, the length of the recordings differs, with most records being 10 seconds in length. To standardise the data length fed into the network, we apply recursive padding and data truncation, which may have a certain impact on the ECG information. Specifically, records less than 10 seconds are padded by reusing the prior segment to achieve a length of 10 seconds, while records longer than 10 seconds are truncated to include only the first 10 seconds of data as the current data.

Federated learning setup. To simulate the real-world *cross-silo* FL setting, we treat each data

Table 7.1: **Architecture of MobileNet_V2 for CIAFR10 classification**. Features used in *FLea* are underlined with $l = 5$.

Block(CNN layers)	#Input	Operator	#Output Channel	#Kernel	#Stride	#Output
$0(1)$	$3 \times 32 \times 32$(image)	conv2d	32	3	1	$32 \times 32 \times 32$
$1(2-5)$	$32 \times 32 \times 32$	conv2d$\times$4	$32, 32, 16, 16$	$1, 3, 1, 1,$	$1, 1, 1, 1$	$\underline{16 \times 32 \times 32}$
$2(6-9)$	$16 \times 32 \times 32$	conv2d$\times$4	$96, 96, 24, 24$	$1, 3, 1, 1$	$1, 1, 1, 1$	$32 \times 32 \times 32$
$3(10-12)$	$32 \times 32 \times 32$	conv2d$\times$3	$144, 144, 24$	$1, 3, 1$	$1, 1, 1$	$24 \times 32 \times 32$
$4(13-14)$	$24 \times 32 \times 32$	conv2d$\times$3	$144, 144, 32$	$1, 3, 1$	$1, 2, 1$	$32 \times 16 \times 16$
$5\&6(15-20)$	$32 \times 16 \times 16$	conv2d$\times$3	$192, 192, 32$	$1, 3, 1$	$1, 1, 1$	$32 \times 16 \times 16$
$7(21-23)$	$32 \times 16 \times 16$	conv2d$\times$3	$192, 192, 64$	$1, 3, 1$	$1, 2, 1$	$64 \times 8 \times 8$
$8, 9, \&10(24-32)$	$64 \times 8 \times 8$	conv2d$\times$3	$384, 384, 64$	$1, 3, 1$	$1, 1, 1$	$64 \times 8 \times 8$
$11(33-36)$	$64 \times 8 \times 8$	conv2d$\times$4	$384, 384, 96, 96$	$1, 3, , 11$	$1, 1, 1, 1$	$9 \times 8 \times 8$
$12\&13(37-42)$	$96 \times 8 \times 8$	conv2d$\times$3	$576, 576, 96$	$1, 3, 1$	$1, 1, 1$	$96 \times 8 \times 8$
$14(43-45)$	$96 \times 8 \times 8$	conv2d$\times$3	$576, 576, 160$	$1, 3, 1$	$1, 2, 1$	$160 \times 4 \times 4$
$15\&16(46-51)$	$160 \times 4 \times 4$	conv2d$\times$3	$960, 960, 160$	$1, 3, 1$	$1, 1, 1$	$160 \times 4 \times 4$
$17(52-54)$	$160 \times 4 \times 4$	conv2d$\times$3	$960, 960, 320$	$1, 3, 1$	$1, 1, 1$	$320 \times 4 \times 4$
$18(55)$	$320 \times 4 \times 4$	conv2d	1280	1	1	$1280 \times 4 \times 4$

source from the above-mentioned CIC2020 database as a data silo, and those institutions cannot exchange data. To evaluate the performance of the FL algorithms, we initially reserve 20% of the data from each resource to create a global testing set, ensuring it covers all 27 classes. Subsequently, we distribute the remaining data to six clients, keeping the original independence.

For *FLea*, we assume that the features from the basic blocks can be shared globally, as illustrated in Figure 7.4. Since the amount of data varies significantly across the six clients, we consider sharing a fixed amount of feature-target pairs for *FLea*: 50, 200, 500, and 1000, respectively (for St. Petersburg, we always share 30 features as it only owns 74 samples in total, and for PTB we sharing up to 300 features). We use the Adam optimiser for local training with an initial learning rate of 10^{-3} and decay it by 2% per communication round until 10^{-5}. The size of the local batch is 64, and we run 1 local epoch per round and 100 communications in total. The best global model during training will be reported for comparison.

7.4.2 Setup for CIFAR10

Data and model. To gain a deeper understanding of our proposed method and to verify its generality across various data distributions, we further evaluate *FLea* via the machine learning benchmark CIFAR10 (Krizhevsky et al., 2009) (introduced in Chapter 2.2.3). We classify images in CIFAR10 using the MobileNet_V2 model that has 18 blocks consisting of multiple convolutional and pooling layers (Sandler et al., 2018). The images are cropped to a size of 32×32, for both training and testing. The architecture of MobileNet_V2 is summarised in Table 7.1.

Federated learning setup. To simulate the decentralised setting, we distribute the training set (comprising 50,000 samples for 10 classes) to 100 and 500 clients for model development

and utilise the testing set (containing 1,000 samples per class) to report the accuracy of the global model. Following strategies from (Zhang et al., 2022), we employ quantity-based heterogeneity ($Quantity(q)$) and distribution-based heterogeneity ($Dirichlet(\mu)$) for data splits. $Quantity(q)$ suggests each client has q classes of data, with the number of samples in each class being the same. For $Dirichlet(\mu)$ split, μ controls the heterogeneity of all classes on each client, whereas a smaller μ suggests a more uneven distribution of data across clients.

We use the Adam optimiser for local training with an initial learning rate of 10^{-3} and decay it by 2% per communication round until 10^{-5}. The size of the local batch is 64, and we run 10 local epochs for 100 clients setting and 5 local epochs for 500 clients setting. Considering the scale of clients, 10% of clients are randomly sampled at each round to participate in local training and model aggregation. We run 100 communications and take the best accuracy as the final result.

7.4.3 Baselines and metrics

We compare *FLea* against *FedAvg*, and then the state-of-the-art *loss-based* methods: i) *FedProx* (Li et al., 2020b), and ii) *FedNTD* (Lee et al., 2022); as well as *data-based* methods: iii) *FedData* (Zhao et al., 2018) and iv) *FedMix* (Yoon et al., 2020). v) *CCVR* (Luo et al., 2021). These methods have been introduced in Chapter 7.2. We adapt their official implementation to our used datasets. All baselines are hyper-parameter optimised to report their best performances. For *FedMix*, we use the augmentation of the average over every 10 data samples. Besides, regarding the *data-based* methods, for the ECG data, we experiment with each client sharing up to 50 samples or features with others, while for the CIFAR10, each client shares 10% of the local data or features from 10% of the local data. Additionally, for the ECG task, we report the performance of the model trained by the centralised data (merged from the six data silos) as a reference.

For the ECG classification task, to measure the performance, we employ the metrics of ***Sensitivity*** and ***Specificity***, and ***Youden's index*** (the same metrics used in the experiment from Chapter 6.4). Here, we report Sensitivity and Specificity using an operating threshold of 0.5, as the output probability ranges from 0 to 1. For Youden's index, we search the threshold on the ROC curve to report the maximum value for comparison. Acknowledging the non-uniform distribution of the 27 classes within the testing set, we opt to present the macro-averaged metrics across these classes, thereby furnishing a more comprehensive evaluation of the FL algorithms regarding health diagnostics accuracy. Specifically, we report ***Macro-Sensitivity*** and ***Macro-Specificity***, and ***Macro-Youden***. Mathematically, Macro-Sensitivity$= \frac{1}{C}\sum_{c=1}^{27}\text{Sensitivity}_c$, where Sensitivity_c is the Sensitivity for class c. The same applies to Macro-Specificity and Macro-Youden. Furthermore, for both the baseline and our proposed methods, we report the 95% Confidence Interval (CI) for all metrics by using bootstrap, as consistently used in the previous

Table 7.2: **Performance comparison for ECG classification.** 95% CIs are reported in brackets. We report the performance for Method(x) with (x) denoting up to x samples or features from each client are shared globally.

	Macro-Sensitivity	**Macro-Specificity**	**Macro-Youden**
Centralised	0.553(0.542-0.567)	0.985(0.979-0.998)	0.558(0.530-0.585)
FedAvg	0.308(0.277-0.325)	0.965(0.954-0.976)	0.293(0.251-0.318)
FedProx	0.314(0.279-0.328)	0.970(0.959-0.982)	0.298(0.256-0.326)
FedNTD	0.367(0.360-0.375)	0.926(0.919-0.934)	0.313(0.279-0.326)
FedData (50)	0.420(0.411-0.432)	0.971(0.968-0.990)	0.411(0.379-0.442)
FedMix (50)	0.361(0.346-0.390)	0.946(0.921-0.973)	0.315(0.307-0.338)
CCVR (50)	0.335(0.319-0.348)	0.953(0.939-0.970)	0.305(0.290-0.318)
FLea (50)	0.383(0.376-0.395)	0.974(0.965-0.987)	0.370(0.361-0.387)
FLea (200)	0.424(0.415-0.437)	0.980(0.971-0.993)	0.415(0.423-0.445)
FLea (500)	0.441(0.436-0.458)	0.980(0.972-0.994)	0.422(0.428-0.461)
FLea (1000)	0.450(0.439-0.460)	0.981(0.974-0.995)	0.440(0.430-0.463)

chapter.

For the CIFAR10 benchmark, since the testing set is balanced across classes, we report the overall ***Accuracy*** for comparison (as formulated in Chapter 3.4). We run all experiments five times with different random seeds to derive the mean and standard deviation for the metrics as well.

7.5 Results on multi-centre ECG data

Now, let us look at the experimental results based on the multi-centre ECG data.

7.5.1 Comparison to baselines

The results are summarised in Table 7.2, which deliveries the following observations:

- **Given the notably disparate class distribution across the six data silos, the performance of *FedAvg* performance degrades markedly compared to the centralised setting.** In particular, *Sensitivity* experiences a relative reduction of 44.3% (from 0.553 to 0.308), suggesting a substantial number of missed cardiovascular diagnoses. This also confirms our motivation that the heterogeneous physiological data distribution across different data centres poses a great challenge to federated learning.

- **The cutting-edge loss-based baseline, *FedNTD*, factors in the local data distribution while retaining the global model's knowledge during local training.** This adaptation effectively enhances *Sensitivity* compared to *FedAvg*, albeit at the expense of a decrease

in *Specificity*. Consequently, its *Youdex's index* exhibits a marginal difference from that of *FedAvg*, *i.e.*, 0.313 versus 0.293. This further indicates that a merely loss-based FL approach is still insufficient to handle the data heterogeneity of real-world physiological data.

- **Compared to the loss-based baselines, the data-based baselines, including *FedData*, *FedMix*, and *CCVR*, are more effective.** They enclose the gap between *FedAvg* and centralised training, confirming our hypothesis that a global proxy can effectively mitigate the local bias caused by the absence of certain classes. Among the three methods, *FedData* performs the best. However, it is important to note that this method requires clients to share raw data and labels, making it infeasible for healthcare applications.

- ***FLea* is superior to data-based baselines.** By sharing the same amount of augmentations (i.e., up to 50 samples or features per client), *FLea* outperforms *FedMix* and *CCVR* with notable performance gain. Specifically, its *Youdex's index* improves *FedMix* relatively by 17.4% (from 0.315 to 0.370) and *CCVR* by 21.3% (from 0.305 to 0.370), respectively. In comparison to *FedData*, *FLea* yields competitive *Sensitivity* and *Specificity* with the privacy of the data better preserved. It is worth noting that feature exposure is not equivalent to privacy leakage, as the features of *FLea* do not leak source data. We will further provide a more comprehensive comparison to *FedData* and *FedMix* in terms of privacy leakage in Chapter 7.6.2.

- **Upon sharing more features (from 50 to 1000) within each training round, the performance enhancement of *FLea* becomes more pronounced:** *Sensitivity* climbs from 0.383 to 0.447, significantly mitigating the performance gap between decentralised training and centralised training. Additionally, it is noteworthy that when more than 500 features are shared from each client, the improvements in both *Sensitivity* and *Specificity* stabilise. This observation implies that sharing 500 features strikes the optimal balance between diagnostic accuracy and training efficiency.

7.5.2 Case study

Let us further explore the performance of *FLea* when sharing up to 500 features. While the 500 features are a small amount in comparison to the vast local dataset residing in the CPSC (6877 samples), CPSC-Extra (3453 samples), PTB-XL (21837 samples), and G12ECG (10344 samples), the brought performance gain in terms of *Sensitivity* is remarkable (*Specificity* for those methods are similar and consistently high so we don't particularly compare): we observe an impressive 43.2% improvement over *FedAvg* and a noteworthy 20.2% enhancement over *FedNTD*.

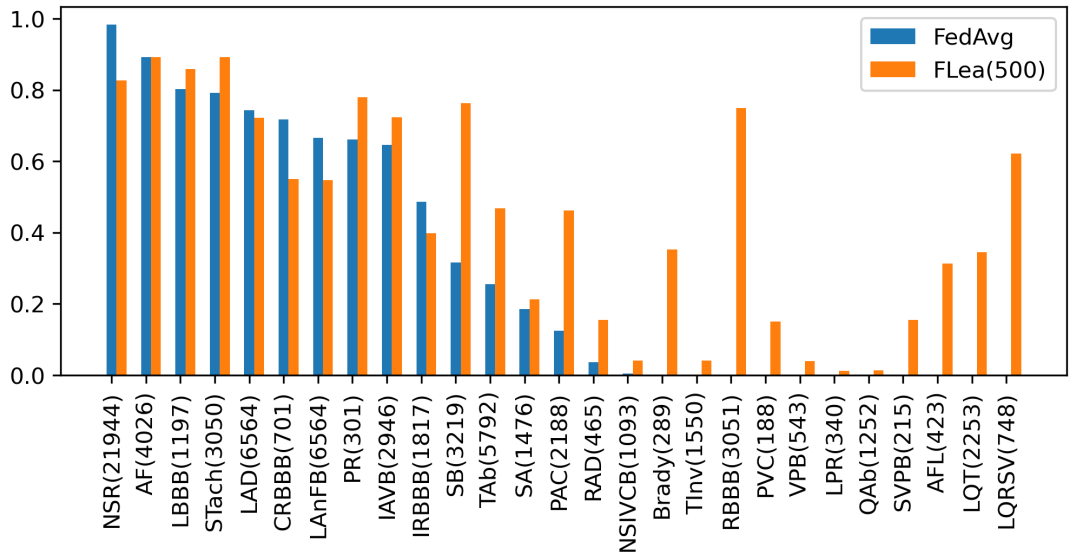
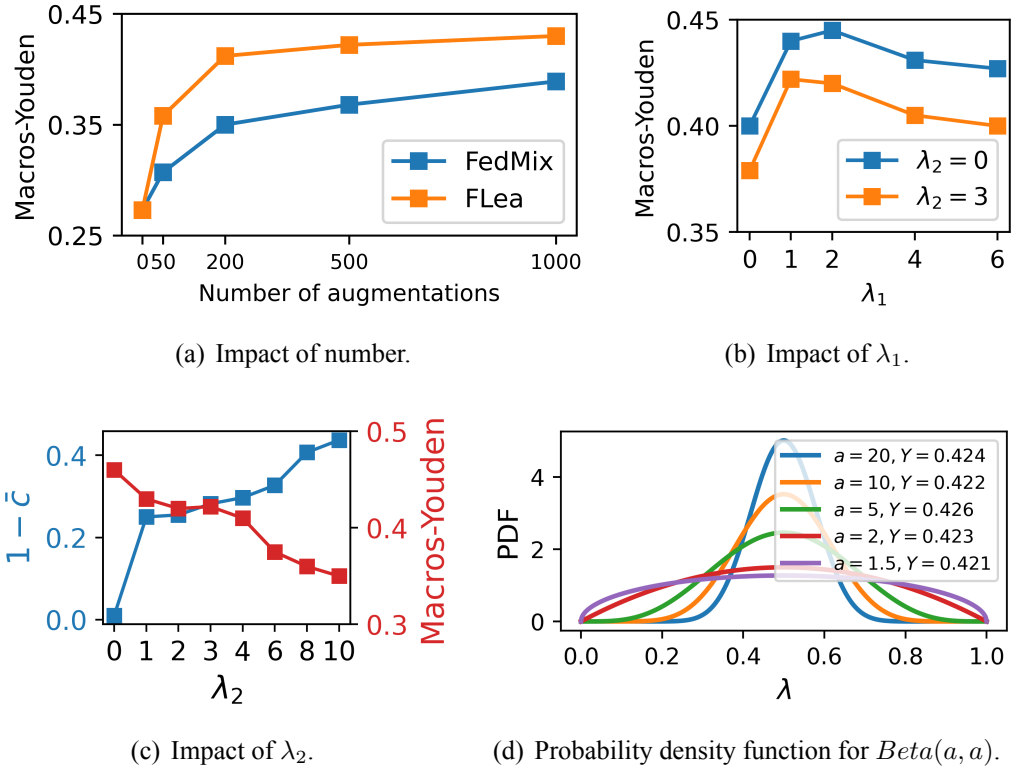


Figure 7.5: **Performance comparison between *FedAvg* and *FLea***. *Sensitivity* values are compared, and diagnoses are ranked based on the *Sensitivity* of FedAvg using the testing set. The x-axis presents the diagnosis abbreviations alongside the total number of samples for each diagnosis enclosed in brackets. The distribution of each diagnosis across clients is illustrated in Figure 7.3.

To further understand how *FLea* improves the *Sensitivity*, we visualise the *Sensitivity* for each class in Figure 7.7. It can be first observed that *FedAvg* struggles to accurately diagnose a majority of abnormalities, yielding a sensitivity lower than 0.5 even for two-thirds of the classes. More interestingly, the model developed by *FedAvg* failed to diagnose those abnormalities not because they are under-represented in the global data distribution. On the contrary, some of them are certain predominant classes such as NSIVCB (1093 samples), RBBB (3051 samples), and LQT (2253 samples). When we look at the data distribution in Figure 7.3, it is easy to figure out that NSIVCB, RBBB, and LQT exhibit pronounced distribution heterogeneity, with the majority of samples concentrated in one or two silos. It is evidently that the data distribution disparity poses a challenge to *FedAvg*. In this regard, *FLea*(500), through leveraging globally shared features to augment the local training set, proves to be remarkably effective in mitigating the performance degradation caused by data disparity. Consequently, it consistently enhances *Sensitivity* across all classes compared to *FedAvg*.

7.5.3 Impact of hyper-parameters

Table 7.2 presents the results for sharing different numbers for features for *FLea*. Following that, Figure 7.6(a) further shows the comparison for *FLea* and *FedMix*. It is evident that a number of 50 can significantly boost the performance while when sharing more, the advantage over *FedMix* still remains. *FedMix* uses the average of data samples over mini-batch to protect data



(a) Impact of number.

(b) Impact of λ_1.

(c) Impact of λ_2.

(d) Probability density function for $Beta(a, a)$.

Figure 7.6: **Hyper-parameter tuning for ECG classification**. In (a), the performance of sharing different numbers of augmentations by *FLea* and *FedMix* are compared. (a) and (b) present the impact of λ_1 and λ_1 in Eq. (7.5), where $\bar{c}$ denotes the averaged correlation between the feature and the original data (refer to Eq. (7.4)) for all rounds. (d) illustrates the Beta distribution with varying a, where the yielded *Macro-Youden* is annotated (shorted as Y).

privacy, while *FLea* leverage the features. These results demonstrate the superiority of *FLea* in boosting the performance.

We then illustrate how we identify the hyper-parameters λ_1 and λ_2 for the loss function (Eq. (7.5)) and a in the $Beta$ distribution for the augmentation (Eq. (7.1)) in Figure 7.6. We first set $\lambda_2 = 0$ (without obfuscating the features) and search the value for λ_2. As shown in Figure 7.6(b), we found that $\lambda_1 > 1$ can improve the performance compared to that without the distilling loss ($\lambda_1 = 0$), but if the weight is too large ($\lambda_1 > 4$) it harms the performance. The pattern is similar with other λ_2, and thus we informally use $\lambda_1 = 1$ for all experiments. With $\lambda_1 = 1$, we further study how λ_2 impacts the trade-off between privacy preservation (reflected by the reduced correlation c) and the feature utility (reflected by the *Macro-Youden*), as shown in Figure 7.6(c). Enlarging λ_2 can significantly enhance privacy protection (referring to the increasing $1 - \bar{c}$) but decreases the final performance. We finally use $\lambda_2 = 3$ when the $\bar{c}$ reduces to about 0.72 while maintaining a strong *Macro-Youden* of about 0.42. We also suggest future applications using $2 \sim 6$ for the trade-off. In Figure 7.6(b), we demonstrate that the final performance is not sensitive to the parameter of the $Beta$ distribution since we always have an expectation of 0.5

Table 7.3: **Overall performance comparison for CIFAR10**. Accuracy is reported as $mean \pm std$ across five runs. The best baseline (excluding *FedData*) under each column is highlighted.

	#Clients: 100 (500 samples per client on average)			**#Clients: 500 (100 samples per client on average)**		
%	$Quantity(3)$	$Dirichlet(0.5)$	$Dirichlet(0.1)$	$Quantity(3)$	$Dirichlet(0.5)$	$Dirichlet(0.1)$
FedAvg	43.55±0.82	50.36±0.89	28.21±1.20	30.25±1.33	32.58±1.09	20.46±2.15
FedProx	44.37±0.89	49.30±1.00	34.66±1.11	31.92±1.45	32.01±1.25	20.86±1.97
FedNTD	53.01±1.23	56.06±0.97	41.48±0.90	39.98±0.97	39.82±0.86	26.78±2.34
FedData	67.60±1.33	72.17±1.34	70.34±1.68	54.64±1.02	56.47±1.22	55.35±1.46
FedMix	52.78±1.99	57.97±1.24	40.68±1.50	44.04±1.53	45.50±1.88	38.13±2.06
CCVR	49.11±0.67	51.21±0.98	34.47±1.35	35.95±1.63	35.02±1.43	24.21±2.67
FLea ($l = 5$)	58.27±0.95	59.63±1.28	43.65±1.47	47.03±1.01	48.86±1.43	44.40±1.23

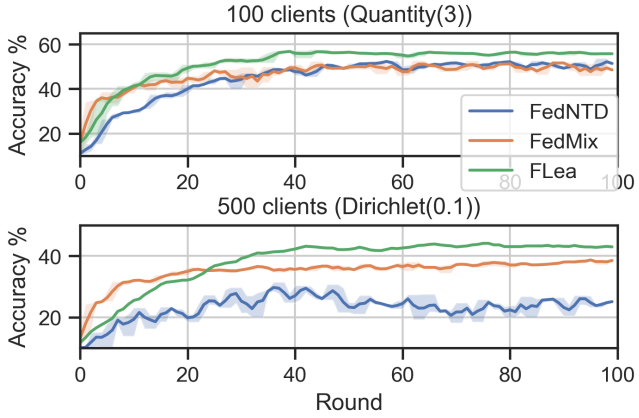


Figure 7.7: **Accuracy of the model in each communication round**. Two examples are given: (a) shows the results for 100 clients with each client having 3 classes of data; (b) shows the results for 500 clients with heavily heterogeneous local data.

for λ.

In summary, in this multi-centre ECG classification task, *FLea* significantly outperforms existing FL methods in terms of diagnostic accuracy. Additionally, *FLea* demonstrates competitive performance compared to centralised training methods and FL approaches involving raw data sharing, all while mitigating the risk of privacy leakage through feature sharing.

7.6 Results on distributed machine learning benchmark data

To assess the generality of our proposed method, *FLea*, in handling data heterogeneity in federated learning, we also conduct extensive experiments using the CIFAR10 data with varying distributions. The results are summarised and discussed below.

7.6.1 Performance comparison

We summarise the overall comparison to baselines for the benchmark data in Table 7.3. From the table, we can observe that *FLea* consistently outperforms the baselines and closes the gap to *FedData* across different data splits. Quantitatively, *FLea* outperforms the start-of-the-art FL method (*FedNTD*) that sharing only model parameters, by up to 17.22% (from 26.78% to 44.40%), and FL methods that share data augmentations (*FedMix*) by up to 6.24% (from 38.13% to 44.04%), while reducing the privacy vulnerability associated with shared data augmentations. Moreover, *FLea* presents more stable performance compared to *FedNTD* and *FedMix*. As shown in Figure 7.7, *FLea* converges after 40 communication rounds, with notably higher averaged accuracy and smaller variance compared to the other two best baselines.

It is noteworthy that *FLea* exhibits even greater superiority when dealing with smaller local data sizes. Specifically, as the number of clients decreases from 100 to 500, resulting in significantly smaller local data sizes (approximately 100 samples) for model training, the performance advantage of *FLea* over *FedNTD* becomes more pronounced. This observation suggests that *FLea* can not only mitigate local drift caused by label distribution heterogeneity across clients but also alleviate local overfitting resulting from limited training data. This is achieved through our novel feature-sharing and augmentation mechanisms.

7.6.2 Privacy protection analysis

In addition to boosting the performance, *FLea* aims to mitigate the privacy leakage associated with feature sharing from two aspects: *i)* defending from data reconstruction attack, and *ii)* preventing the sensitive context information from being identified. We now demonstrate *FLea* is more privacy-preserving than *FedMix* and *FedData* as follows.

First of all, we visualise the comparison among a raw data sample, the augmentation used by *FedMix* and the feature used by *FLea* in Figure 7.8. It is worth noting that feature exposure is not equivalent to privacy leakage, as the features of *FLea* do not leak source data. To quantify the privacy exposure risk, we set two privacy attacks, *i.e.*, data reconstruction and context identification, by using $Quantity(3)$ data splits and $K = 100$ as an example for studying. Since in other settings either the label is more skewed or the local data is more scarce, a privacy attack can hardly be more effective than in this setting. This is to present the attack defending for the most vulnerable case. As the correlation between the features and the data is continuously reduced by our de-correlation loss during the entire training procedure, we report the results for the $c = 0.65$ (the 1^{st} round) and $c = 0.40$ (the 10^{th} round) for reference (c presents the correlation between the feature and the data in a certain round).

Data reconstruction. We first implemented a data reconstruction attacker to test whether the original data can be recovered from the shared features. Following the approach described

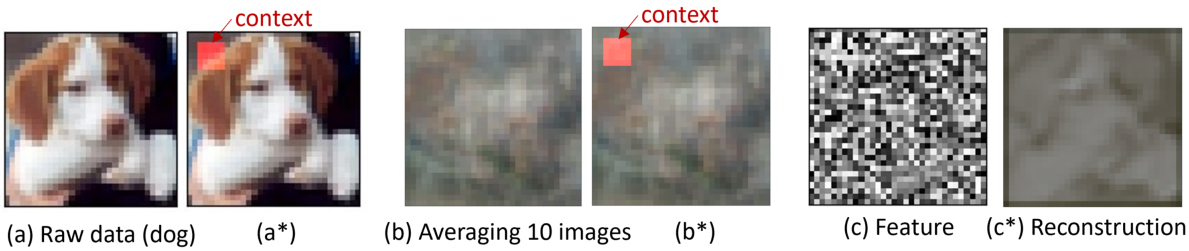
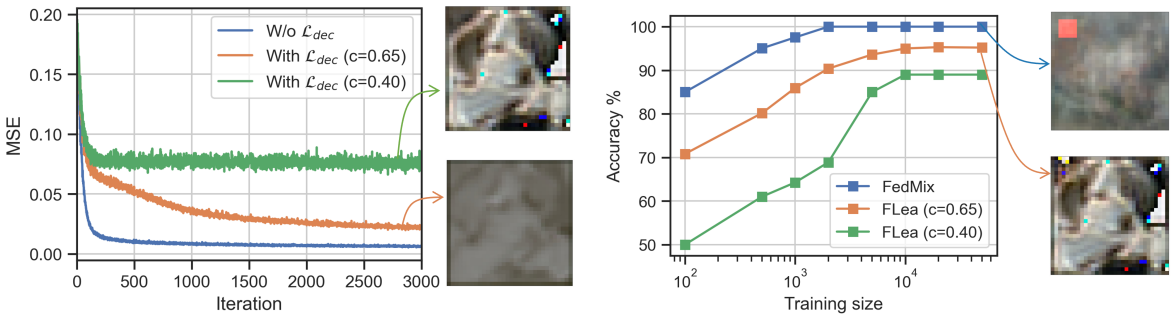


Figure 7.8: **Visualisation for data and data augmentations.** (b) is the average of a batch of samples like (a), but if the local data contains individual context information (*e.g.,* (a*)), averaging over those samples cannot protect such information (*e.g.,* (b*)). (c) shows a feature of (a*) and (c*) shows its reconstruction. (b) is used by *FedMix* and (c) is used by *FLea*. From (a) to (c), the privacy vulnerability is reduced.



(a) Reconstruction training error. (b) Context detection accuracy.

Figure 7.9: **The effectiveness privacy protection.** c is short for the correlation as defined in Eq. 7.4. We show the reconstruction and context detection performance for $c = 0.65$ (the 1^{st} round) and $c = 0.40$ (the 10^{th} round).

.

in (Dosovitskiy and Brox, 2016), the attacker builds a decoder model for the purpose of reconstruction. Specifically, the decoder architecture, designed to match the MobileNet_V2 architecture, comprises four *conv2d* layers (refer to Table 7.4) to reconstruct the original data from the provided features. For visualisation purposes, the CIFAR10 images are cropped to a size of 32×32 pixels without any normalisation. The decoder takes the features extracted from the global model as input and generates a reconstructed image, which serves as the basis for calculating the mean squared error (MSE).

To train the decoder, we utilise the entire CIFAR10 training set, conducting training for 20 epochs and employing a learning rate of 0.001. This approach allowed us to evaluate the fidelity of the reconstructed data and compare it with the original input, providing insights into the effectiveness of our proposed feature interpolation method.

We use the testing set and the target global model ($c = 0.65$ and $c = 0.40$) to extract features for reconstruction. Figure 7.9(a) shows the training process for the decoder model, while the

Table 7.4: **Architecture of decoder of MobileNet_V2**. The input feature with a dimension of $16 \times 32 \times 32$ is fed into this model to generate an image with a dimension of $3 \times 32 \times 32$.

Layer Index	#Input	Operator	#Output Channel	#Kernel	#Stride	#Output
1	$16 \times 32 \times 32$ (Feature)	conv2d	32	1	1	$32 \times 32 \times 32$
2	$32 \times 32 \times 32$	ConvTranspose2d	32	3	2	$32 \times 64 \times 64$
3	$32 \times 64 \times 64$	conv2d	32	3	2	$32 \times 32 \times 32$
4	$32 \times 32 \times 32$	conv2d	3	1	1	$3 \times 32 \times 32$ (Data)

training MSE is presented in the curves while the exampled images are from the testing set. Firstly, it is clear that without obfuscating the features, as shown in the $\mathcal{L}_{dec}$ group, the training MSE can be progressively reduced to about 0.0. However, with the $\mathcal{L}_{dec}$ ($c = 0.4$ or $c = 0.65$), the training loss stops to further reduce after several iterations. This suggests that the training data can not be well recovered from the features. More straightforwardly, let us look at the recovered image for the feature extracted in Figure 7.8(a). As displayed by the two images in Figure 7.9(a): i) for $c = 0.65$ the sensitive attributes are removed (*e.g.,* the colour of the dog), and ii) for $c = 0.4$, only some textures are recovered, while other information is not. Overall, with $\mathcal{L}_{dec}$, the correlation between data and features is reduced, preventing the image from being reconstructed from the feature. This demonstrates the effectiveness of *FLea* in mitigating the privacy exposure associated with feature sharing.

Context information identification. In the second attack, we aim to test whether privacy-related context information can be released from the shared features. Privacy-related context refers to some personal information embedded in the data. Considering an application where the client's phone has a camera sensor problem so that each photo has a spot (see Figure 7.8(a*))), or the client lives in a busy neighbourhood and thus all audio clips have a constant background score. Releasing such context information may lead to personal identity leakage. Averaging over a batch of samples will not protect such context information, as shown in Figure 7.8(b*).

We assume that the attacker explicitly knows the context information and thus can generate large amounts of negative (clear data) and positive (clear data with context marker) pairs to train a context classifier (which is very challenging and unrealistic but this is for the sake of testing). Real-world attacks will be far more challenging than our simulations. The context identification attacker is interested in finding out if a given feature f, is from the source data with a specific context or not. We simulate the context information by adding a colour square to the image (to mimic the camera broken), as illustrated in Figure 7.8. We use a binary classifier consisting of four linear layers to classify the flattened features or images. To train the classifier, we add the context marker to half of the training set to simulate the contexts, while the rest is clear data. Similar to the training data, we add the context marker to half of the testing set to evaluate the context identification performance.

We present the identification accuracy achieved with varying amounts of training data in Figure 7.9(b), comparing *FedMix* and *FLea*. From the results in Figure 7.9(b), it is initially observed that, with an equal number of available training samples, the identification accuracy for *FLea* is consistently 20∼40% lower than that for *FedMix*. Conversely, to achieve a comparable level of accuracy, *FLea* requires a larger number of training samples than *FedMix*. This underscores *FLea*'s ability to better safeguard contextual privacy. These findings collectively indicate that *FLea* enhances the complexity of potential attacks, thereby providing improved protection for contextual privacy compared to *FedMix*.

All the above results lead to the conclusion that by reducing and mitigating the correlation between the features and source data, *FLea* safely protect the privacy associated with feature sharing while achieving favourable performance gain in addressing various data distribution disparities in FL.

In summary, the extensive experimental results presented above showcase the superiority of *FLea* from various perspectives. *FLea* is an effective framework for health institutes to collaboratively develop health diagnostic models without exchanging their collected physiological datasets. Our work holds significant promise in effectively addressing the challenging dilemma that isolated data is insufficient while aggregating data poses a privacy risk, and thus paves the way for trustworthy machine learning in healthcare applications.

7.7 Discussion and conclusions

This chapter explored an orthogonal dimension from the study presented in Chapter 6: the *cross-silo* FL applied to distributed physiological data. In contrast to *cross-device* FL, the *cross-silo* approach involves each silo containing a larger volume of data, albeit with data spanning multiple disease categories. Consequently, a pronounced heterogeneity emerges in both data volume and class distribution across these silos. To address this intricate challenge and facilitate collaboration among health institutions without data exchange, we introduced *FLea*. This innovative algorithm, based on feature sharing, enables high-performing FL for such decentralised physiological data. Our approach was thoroughly validated through a combination of empirical assessments using both real-world and benchmark datasets.

We recognise a limitation within this work: the local label distribution may be disclosed to the server during feature sharing. This could raise concerns, particularly when the client-owned classes are sensitive, such as individual disease diagnoses. This drawback renders *FLea* unfeasible for cross-device scenarios wherein label distribution cannot be enclosed. While it is worth noting that many existing FL methods also explicitly reveal label distribution to the server (Luo et al., 2021; Tan et al., 2022), and for other methods, label distribution can potentially be in-

ferred from the gradients (Dang et al., 2021; Wei et al., 2020; Wainakh et al., 2022), we are committed to enhancing *FLea* by sharing label-agnostic features in the future. This strategic modification will render the algorithm adaptable to a broader array of scenarios while retaining a robust privacy protection capability.

In both Chapter 6 and 7, our primary focus was on addressing the heterogeneity in local data distribution, specifically concerning label distribution. While this represents a key challenge in real-world federated learning applications, there are instances where semantic shifts can occur alongside label skew. Semantic shifts occur when data collected by mobile devices or institutions exhibit domain-specific differences due to variations in devices and protocols (Li et al., 2020c; Chen et al., 2023b). This has inspired us to delve further into this issue as an extension of our work.

Chapter 8

Conclusions and future directions

The future of work lies in the collaboration between humans and AI.

- Demis Hassabis

co-founder and CEO of DeepMind

8.1 Summary of contributions

In the introduction of this thesis, we highlighted the challenges arising from limited and imbalanced physiological data, overconfidence in deep learning, and the privacy protection requirements for health-related machine learning research. Despite numerous efforts, including those within the broader machine learning literature, these challenges remained inadequately addressed. The primary objective of this thesis was to address these issues and make a meaningful contribution to the development of reliable deep-learning methodologies for real-world health diagnostic applications. We now delve into a specific reflection on the research questions introduced in Chapter 1 and provide a concise summary of the major contributions made in this thesis.

***Research Question 1**: How can we mitigate the bias and calibrate the confidence of predictions when training models for health screening with limited and imbalanced physiological data?*

Contribution 1: The accuracy and the reliability of automated diagnostics are equally crucial for real-world medical applications. However, deep learning is susceptible to bias and overconfidence when developing models using limited and imbalanced physiological data.

To enhance the reliability of deep learning in the health domain, Chapter 4 introduced a novel

data-balanced ensemble learning method. This approach involves re-sampling imbalanced data to create class-balanced subsets for training model ensembles. The effectiveness of this design was validated in the context of a physiological audio-based COVID-19 screening task. Our approach significantly improved screening accuracy by mitigating overconfident predictions of a single deep-learning model. Furthermore, by utilising quantified model uncertainty from the ensembles for selective prediction, our methods demonstrated an additional screening accuracy boost of 17.6%.

***Research Question 2**: How can we develop high-performing models with efficient uncertainty estimation for health diagnostics, given multi-class imbalanced physiological data?*

Contribution 2: Mobile health applications require not only effective uncertainty quantification but also efficiency, given the limited computational resources of devices. Evidential deep learning (EDL) represents the state-of-the-art in efficient uncertainty quantification, capable of estimating predictive uncertainty through a single model and a single forward pass. However, this approach is susceptible to class imbalance. In order to make EDL effective for health diagnostics applications, where physiological data often exhibit imbalance, Chapter 5 introduced a class-balanced EDL method with two novel mechanisms: *i)* a class-level pooling loss to mitigate bias in classification evidence, and *ii)* a learnable prior, regulated by the class distribution, to facilitate learning for minority classes.

The superiority of our method was demonstrated through extensive experiments using three physiological datasets and one machine learning benchmark with varying degrees of class imbalance. Results suggested that our method not only enhanced diagnostic accuracy but also reduced overconfident predictions by up to 43% compared to other uncertainty-aware baselines, while keeping as efficient as the traditional *Softmax*-based method. We also introduced the use of uncertainty measurements for misdiagnosis identification and out-of-training-distribution detection. Consequently, our method outperforms state-of-the-art methods in these applications by up to 16.1%. Our work paves the way for uncertainty-aware mobile health applications.

***Research Question 3**: How can we train deep learning-driven health screening models using only decentralised and imbalanced physiological data stored on mobile devices?*

Contribution 3: Privacy concerns pose a significant obstacle to health-related data collection for machine learning research. In Chapter 6, we delved into the realm of *cross-device* federated learning, aiming to facilitate model training using physiological data distributed across mobile devices. Acknowledging the local and global class imbalances inherent in decentralised physiological data, we proposed a weighted federated learning aggregation method. This aggregation

of local models is guided by the model loss, addressing biases induced by imbalanced physiological data.

The effectiveness of this method was validated through the physiological audio-based COVID-19 screening task, assuming that physiological audio samples remain where they were collected. Results demonstrated the superior performance of our method, comparable to centralised training outcomes. Our study opens the door to privacy-preserving mobile health research.

> ***Research Question 4***: *How can we develop high-performing models for health diagnostics using physiological data distributed in multiple places and with heterogeneous distributions?*

Contribution 4: For distributed physiological data with heterogeneous health condition distributions, Chapter 7 introduced a novel *cross-silo* federated learning approach. This approach enables multiple data holders to collaboratively develop a health diagnostic model without exchanging their raw data. To address the issue of data heterogeneity, our method leveraged globally shared features as an augmentation to enhance local training, reducing the discrepancy between local and global models.

The effectiveness of our method was first validated using real-world multi-centre ECG data. The results demonstrated its capability to develop a high-performing cardiac arrhythmia detection model without centralising the ECG data. Additionally, extensive experiments were conducted on machine learning benchmark data. The results not only indicated the superior performance of our method compared to state-of-the-art federated learning baselines but also suggested its privacy-preserving advantages over existing data-sharing-based federated learning counterparts. Therefore, our work facilitates privacy-preserving machine learning research in the health domain.

Conclusions. Beyond the numerical improvements over the compared physiological data modelling methods, my studies have led to two significant high-level breakthroughs in this field: *i)* the transformation of overconfident deep learning predictions into calibrated health diagnoses; and *ii)* the shift in deep learning research from reliance on centralised physiological data to decentralised data. These advancements underscore the importance of accuracy and reliability in health-related machine learning applications, emphasising the need for models that not only perform well statistically but also align closely with the complexities and variability inherent in real-world healthcare data.

The first breakthrough, the transformation of overconfident deep learning predictions into calibrated health diagnoses, addresses a critical challenge in healthcare applications of deep learning: the tendency of these models to make predictions with unwarranted certainty. By im-

plementing mechanisms for uncertainty quantification, my researches provide a robust framework for interpreting deep learning outputs in a manner that is both scientifically grounded and clinically relevant. Firstly, it enhances patient safety by reducing the likelihood of erroneous decisions based on overconfident predictions. When models convey uncertainty in their diagnoses, clinicians can make more informed choices about when to rely on AI-supported insights and when to seek additional information or alternative diagnostic paths. This is particularly crucial in high-stakes medical scenarios where the cost of a mistake is substantial. Moreover, incorporating uncertainty quantification into health diagnostics encourages a more collaborative relationship between AI systems and healthcare professionals. Instead of viewing AI as a definitive authority, clinicians can interpret AI-generated diagnoses as a consultative tool that offers insights while acknowledging its limitations. This paradigm shift fosters a multidisciplinary approach to patient care, where technology complements rather than replaces human expertise.

The significance of the second breakthrough, the shift from reliance on centralised physiological data to decentralised data in deep learning research, marks a paradigm shift with profound implications for the future of healthcare. This breakthrough signifies several key advancements and opportunities. Firstly, it will broaden data accessibility. Decentralising data sources democratises the access to and availability of health-related data. It enables the collection of a wider array of data types from diverse populations across different geographical locations, contributing to a more inclusive understanding of health and disease. This broadening of data sources is critical for developing models that are more representative of the global population, thus improving the unreliability and applicability of deep learning models in healthcare. Secondly, it can enhance research collaboration. The shift towards decentralised data encourages collaboration among researchers, clinicians, and data scientists from around the world. By sharing and analysing decentralised datasets, the research community can uncover novel insights, identify trends, and validate findings across various populations and conditions, accelerating the pace of innovation in medical research.

8.2 Discussion and implications

This thesis explores three directions in physiological data-driven personal health prediction, devoting significant effort to addressing the issues of class imbalance, model overconfidence, and data privacy. We acknowledge that these problems are widely investigated within the broader field of machine learning. However, addressing them specifically for physiological data remains under-explored. This area requires careful experimental design, from data preparation to metric utilisation. A few takeaways are summarised as follows.

1) Preparation of physiological data demands considerable effort, and transparent data sharing is beneficial for the research community. As an example, we discuss the use of COVID-19

Sounds data in Chapters 4 and 6. This dataset was crowd-sourced from mobile app users, meaning the data was collected in uncontrolled environments. Consequently, data quality was variable; for instance, background TV noise is sometimes audible in cough recordings. Cleaning this data required significant effort before it could be used for research. Furthermore, when deploying machine learning algorithms, the division of data into training, validation, and testing sets presents additional challenges. We consistently employed a user-independent split and carefully balanced demographics to avoid bias. While such considerations are crucial for human-centric studies and are a staple in clinical trials, they are often overlooked in much of machine learning research. For comparison, the machine learning community frequently uses well-curated datasets like CIFAR10, which are randomly split. The ECG and skin lesion data utilised in Chapters 5 and 7 had been cleaned and processed by previous researchers. We are grateful for their open-source contributions. As a form of giving back to the community, we have made the prepared COVID-19 Sounds data publicly available for research purposes.

2) Effective algorithms for physiological data typically require the integration of multiple techniques, embodying a multidisciplinary approach. Given the complexity of various types of physiological data, especially signals, a combination of signal processing and machine learning methods is necessary. For physiological audio, such as respiratory and heart sounds, spectrograms play a crucial role in representing the model input. For other signals like PPG, ECG, EEG, and IMU signals, signal processing techniques are essential for pre-processing the data and reducing noise before feeding it into deep learning models. Techniques such as filtering, normalisation, and signal enhancement can significantly improve the quality of input data, facilitating the learning process for deep learning models by highlighting underlying patterns and minimising confusion caused by noise and irrelevant variations.

There is a longstanding debate over whether signal processing remains necessary in the era of AI, and in my view, the two are not mutually exclusive. On one hand, signal processing can enhance resource efficiency. Deep learning models, particularly those capable of processing raw data directly, often require significant computational resources. Signal processing can simplify the data or extract compact representations, thus reducing the computational demands on the deep learning model. This aspect is particularly crucial for deploying models on devices with limited processing power, such as smartphones and embedded systems. On the other hand, signal processing techniques can provide insights into the nature of the data and the underlying physical phenomena. By analysing signals across different domains (*e.g.*, time, frequency, and spatial), researchers and engineers can uncover insights not readily apparent from raw data or the opaque layers of deep learning models, leading to improved interpretability, an area where current deep learning models often fall short.

Even within the realm of machine learning algorithms themselves, it is essential to employ multiple techniques in tandem. For instance, in Chapter 4, we implemented a combination of data

augmentation, transfer learning, and ensemble learning techniques within a single framework. This approach enabled us to develop an uncertainty-aware COVID-19 detection model using limited and imbalanced audio data. Similarly, in Chapter 5, we integrated transfer learning with federated learning to create a COVID-19 detection model without the need to aggregate audio data on a central server. While machine learning research often delves deeply into a single dimension, modelling physiological data demands that researchers possess a comprehensive understanding of a wide array of machine learning techniques.

3) Training models using decentralised physiological data still faces many real-world challenges. In this thesis, although we utilised various physiological data modalities, they were modelled separately, meaning we modelled one variable at a time. In Chapters 6 and 7, we discussed the feasibility of training a single health diagnostics model using decentralised physiological data, which also presented single-variable-based problems. Specifically, for the audio and ECG data we examined, we addressed the issue of label distribution heterogeneity rather than input heterogeneity. In clinical trials, doctors typically make diagnostic decisions using multiple examinations, involving multi-variables. When decentralised data are employed for model training, it's challenging to ensure that different mobile devices or hospitals can collect the same set of physiological data. Under these circumstances, the effectiveness of federated learning methods becomes questionable. This opens up many avenues for future work.

4) Evaluation metrics and comparison should be both clear and equitable. In much of the machine learning literature, *Accuracy* (the ratio of correctly predicted instances to total samples) is often emphasised as the paramount metric. However, in the context of health diagnostics, the significance of different metrics cannot be overstated, and the optimal choice heavily relies on the specific application scenarios. The selection of appropriate evaluation metrics is crucial not only for validating the effectiveness of models but also for securing trust from medical professionals. Accuracy falls short, especially in cases where the dataset is imbalanced, as it may not accurately reflect the model's performance in predicting less frequent outcomes. Medical practitioners often show a preference for the *AUC-ROC* of a model, as it provides a comprehensive overview of the model's ability to distinguish between two classes. However, the interpretation of AUC-ROC can be non-intuitive, necessitating the additional reporting of Sensitivity (true positive rate) and Specificity (true negative rate) to provide clearer insights. Various strategies exist for selecting an operating point on the ROC curve to report these metrics, including minimising false positives to enhance specificity, minimising false negatives to improve sensitivity, or finding an equilibrium between the two. Whichever criterion is used for the proposed method and for the baseline, it should be clearly defined. This nuanced approach ensures a more detailed and transparent assessment of model performance, particularly in critical fields such as healthcare diagnostics, where the stakes are inherently high.

5) Mobile health represents a promising avenue for ubiquitous health monitoring. Significantly,

half of the datasets utilised in this thesis (namely, COVID-19 Sounds data, ECG5000 data, and HAM1000 skin image data) were all collected using mobile devices. This underscores the considerable potential and illustrates the burgeoning interest within the research community: Mobile health is poised to shape the future. Reports indicate that nine in ten people in the UK own a smartphone (Taylor, 2023), and a fifth of American adults regularly utilise a smartwatch or wearable fitness tracker (VOGELS, 2020). Such widespread usage facilitates the continuous collection of individual physiological data, encompassing vital signs, activities, and responses (Inbamani et al., 2022). The expanding market for mobile health devices may soon make healthcare accessible anytime and anywhere.

It could be argued that sensors in mobile devices may not provide the optimal means for health diagnostics. For instance, clinical settings traditionally rely on ECGs and stethoscopes, whereas wearables employ PPG and microphones as alternatives. While it's not certain that mobile devices can meet medical standards on an individual instance basis, the advancement in sensing technologies is undeniable. More crucially, mobile devices offer the potential for long-term health monitoring, leading to the accumulation of longitudinal data. Although individual measurements may not always be optimal, analysing physiological dynamics over time can offer insights into disease progression. This presents a unique advantage of mobile health over traditional clinical trials.

8.3 Future research directions

As a direct extension of the outcomes of the work that has been conducted in this thesis, the following areas of future work are worth exploring.

8.3.1 Is the model fair? Unbiased deep learning for health diagnostics

If a deep learning model unintentionally introduces biases, it may fail to capture the proper relationship between features and the target outcome. This is particularly concerning in sensitive domains like healthcare: it is vital to ensure that these deep learning technologies do not reflect or exacerbate any unwanted or discriminatory biases that may be present in the data (Yang et al., 2023a). In this thesis, we paid considerable attention to class imbalance, a common factor in physiological data that can lead to biased diagnoses. However, other factors, such as demographics, entities, and socio-economics, could also introduce bias into the model. For instance, in our work where audio recordings were utilised to detect COVID-19, experimental results revealed that language can introduce a shortcut from the input to the prediction because the prevalence of COVID-19 varies with languages, despite language not being a relevant feature for COVID-19 detection (Han et al., 2021a). Mitigating algorithmic biases in healthcare is crucial to enhance the robustness of the system when deployed across diverse populations.

Various methods have emerged to address bias, or mitigate the spurious correlations within the data, broadly categorised into two primary types. The first category entails distributionally robust optimisation, aiming to ensure consistent performance across predefined groups delineated by demographics or geographical factors (Sagawa et al., 2019; Levy et al., 2020; Yang et al., 2023a). For instance, the GroupDRO method was devised to minimise the worst-case training loss, guaranteeing a performance floor for underrepresented groups (Sagawa et al., 2019). In a recent study (Yang et al., 2023a), a reinforcement learning approach was introduced to counter bias across various subgroups. The learning reward function in this case encourages a robust classifier and uniform performance for sensitive attributes simultaneously. The second category employs adversarial training, incorporating an additional adversary module to recognise and alleviate biases (Han et al., 2021b; Rajotte et al., 2021). For instance, Yang *et al.* employed clinical features to diagnose COVID-19, employing an adversary module designed to identify and mitigate site-specific (hospital) and demographic (ethnicity) biases. This intervention resulted in enhanced outcome fairness (Yang et al., 2023b).

While emerging research has emphasised the necessity of unbiased deep learning, existing solutions often necessitate pre-defined groups that themselves could be associated with bias. Beyond these predefined groups, recognising potential confounding factors remains challenging. Here, Explainable AI (XAI) can play a crucial role in reducing bias in deep learning-driven health diagnostics by providing insights into the model's decision making process (Saraswat et al., 2022; Chaddad et al., 2023). XAI techniques allow us to analyse the model's behaviour, identify instances of biased patterns, and pinpoint specific features or data points contributing to biased predictions. Understanding the root causes of bias enables targeted interventions to address these issues effectively.

We would like to highlight a few representative XAI techniques tat can be applied to physiological data. The first one is feature importance method, including *SHAP* (SHapley Additive exPlanations) and *LIME* (Local Interpretable Model-agnostic Explanations). SHAP values explain the prediction of an instance by computing the contribution of each feature to the prediction. SHAP is model-agnostic and provides detailed insights into model behaviour, making it highly valuable for understanding how different physiological parameters influence health outcomes. LIME explains predictions by approximating the local decision boundary around an instance. It is particularly useful for explaining predictions of complex models in an interpretable manner by highlighting which features were most influential for specific predictions. The second is the *attention mechanism*, especially for deep learning. Attention Layers in Neural Networks: In deep learning, attention mechanisms can highlight parts of the input data (*e.g.*, segments of an ECG signal or regions in audio spectrogram) that the model pays more attention to when making a prediction. This method is model-specific and helps in understanding which aspects of the physiological data are deemed important by the model. Another one is visualisation technique.

For image-based physiological data (*e.g.*, dermatological images or radiographs), *saliency maps* can highlight regions of the image that significantly influence the model's predictions. This approach helps in identifying visual patterns or anomalies that are predictive of certain conditions. *T-SNE* (T-Distributed Stochastic Neighbor Embedding) can also be applied. It can reduce high-dimensional data (like high-resolution time series or the activation from a deep neural network) to lower-dimensional spaces for visualisation. This can help in understanding the data distribution and how different physiological states are separated in the model's representation space.

Implementing these XAI methods requires careful consideration of the specific healthcare application, the type of physiological data, and the needs of the end-users, typically healthcare professionals or patients. The goal is to enhance transparency, trust, and actionable insights, thereby facilitating better clinical decision-making and patient care.

8.3.2 Certain or not? Benefits of uncertainty for health applications

As the popularity of deep learning continues to rise, it becomes increasingly crucial to establish the reliability of deep learning models for their effective utilisation in healthcare and well-being. Uncertainty quantification is of ever increasing importance in this aspect. This significance is evident in a recent publication in *Nature Medicine* (Dvijotham et al., 2023), where the proper use of confidence scores from a deep learning model for breast cancer screening resulted in a 25% reduction in false positives at the same false-negative rate, significantly reducing clinician workload by 66%. In Chapter 4.5.2, we also demonstrated how high-quality uncertainty estimates can significantly enhance the system's performance and reliability.

However, despite the numerous proposed uncertainty quantification methods, most of the existing work relies on the assumption that uncertainty-aware deep neural networks can effectively model the underlying training data distribution, either explicitly or implicitly. However, a critical issue arises during training, where the focus remains primarily on utilising in-distribution data while neglecting out-of-distribution data, particularly 'uncertain samples', which are not defined and utilised for training. Consequently, this raises doubts about the reliability of current uncertainty estimates.

Unlike machine learning models that merely learn patterns from given categorical data, clinicians often approach decision-making with varying degrees of confidence (in other words, doubt). Difficult cases may elicit differing opinions among doctors, prompting the need for consultations. In light of this, we propose that when training uncertainty quantification models, we should incorporate the diagnostic difficulty as assessed by multiple clinicians to guide the model's learning process. Recently, we have observed a trend of learning with soft labels, wherein the distribution of labels from multiple annotators is considered (Sridhar et al., 2021; Collins et al., 2022; Han et al., 2017; Raghu et al., 2019). This approach provides stronger

supervision for uncertainty estimation. In the future, it is valuable to apply *soft-label* learning techniques to physiological data to enhance the quality of uncertainty estimates. Nevertheless, acquiring multiple annotations from clinicians can be costly, making it essential to address the challenge of reducing clinicians' workload while ensuring effective learning.

Other the other hand, in this thesis, we presented our contributions to uncertainty-aware health diagnostics and health model learning with decentralised data, subsequently. These two directions of research aim to tackle safety and data privacy concerns associated with health applications. Empirically, the proposed uncertainty estimation method shows promise under the federated learning framework. However, it is important to acknowledge that various realistic factors, such as data heterogeneity across different stakeholders, can harm the quality of uncertainty.

In a recent research (Zhang et al., 2023d), a preliminary study was conducted, revealing a decline in the quality of uncertainty estimates in the decentralised setting compared to the centralised setting. This degradation can be attributed to the variations in data collected from different hospitals or health monitoring devices, which encompass diverse technologies, patient demographics, and disease prevalence. The complexity of data heterogeneity poses a significant challenge when attempting to develop a single global model that performs well on all clients. Personalised federated learning, which involves training tailored models for each client, has shown promise in enhancing uncertainty estimation. In the future, building end-to-end systems that integrate uncertainty quantification into the federated learning framework and devising novel techniques to address data heterogeneity will remain crucial for health applications.

As decentralised machine learning continues to evolve, developing robust, scalable, and interpretable methods for uncertainty quantification will be key to its success, particularly in healthcare applications. This endeavour requires not only methodological advancements but also careful consideration of ethical issues, especially concerning privacy and data governance. The dynamic and distributed nature of these systems presents a fertile ground for research and innovation, promising significant impacts across various domains where machine learning is applied. Despite its importance, numerous challenges remain. Firstly, data heterogeneity poses a significant challenge, as decentralised systems often process data that is not uniformly distributed across nodes. It is crucial to quantify uncertainty accurately in these settings to ensure reliable predictions, despite variations in data quality and quantity. Secondly, communication constraints due to bandwidth or privacy concerns may limit constant interaction between nodes (*e.g.*, devices, servers), necessitating efficient methods to estimate and communicate uncertainty. Lastly, model heterogeneity introduces additional complexity, as nodes may employ different model architectures based on their computational capacities. Overall, the unique characteristics of these systems offer extensive opportunities for groundbreaking research and significant advancements in machine learning applications.

8.3.3 Many or one? Foundation models for physiological data

In Chapter 4.4 and 6.4, we employed the VGGish model pre-trained on a public non-physiological audio database to process physiological audio data. This decision was informed by previous findings demonstrating performance enhancements resulting from pre-training (Xia et al., 2021d). However, considering the presence of other modalities such as ECG and EEG, finding matched data for pre-training these models may not always be feasible. This indicates the need to develop a general model capable of accommodating multiple physiological data modalities and easily adapting to various downstream applications.

The recent rise of Large Language Models (LLMs) has garnered extensive attention due to their outstanding performance across various tasks (Zhao et al., 2023; Brown et al., 2020b; Zhang et al., 2023b). The inherent capacity of LLMs to capture knowledge and concepts has significantly improved with the scaling of model size and the exponential increase in training samples. As a result, these models present significant opportunities for diverse downstream tasks, especially in fields like healthcare, where labelled data is often scarce for model development. The prospect of adapting pre-trained LLMs for processing physiological data holds great promise. The feasibility of applying LLMs to physiological time series data arises from two aspects: 1) LLMs are trained with medical knowledge and thus are capable of understanding physiological dynamics from the signals; 2) LLMs, being pre-trained with an auto-regressive objective, can capture sequential patterns, which are prevalent in physiological time series.

To date, a few preliminary studies have explored this direction. For instance, Liu *et al*. pointed out that existing LLMs, while limited to language models alone, can function as effective few-shot health learners (Liu et al., 2023b). They demonstrated that by fine-tuning the model with physiological and behavioural time-series data, the model could make meaningful inferences on numerous health tasks, encompassing clinical and wellness contexts. These tasks included cardiac signal analysis, physical activity recognition, metabolic calculation (*e.g.*, calories burned), and estimation of stress reports and mental health screeners. Similarly, Zhang *et al*. proposed a unified and versatile Biomedical Generative Pre-trained Transformer (BiomedGPT) model (Zhang et al., 2023a). This model leverages self-supervision on large and diverse datasets to accept multi-modal inputs and perform a range of downstream tasks. The experiments showed that BiomedGPT delivered extensive and comprehensive representations of biomedical data.

The pre-trained LLMs can be fine-tuned for physiological data to enhance performance in healthcare applications (Liu et al., 2023a). The main fine-tuning methods include: *i) Transfer learning*. This involves adjusting the pre-trained model on a new, but related, task. The model retains its learned knowledge, which is then refined using a smaller dataset specific to the healthcare task at hand, such as patient diagnosis or treatment recommendation based on physiological data. *ii) Few-shot learning*. Leveraging the ability of LLMs to perform tasks with minimal examples,

few-shot learning fine-tunes the model on a very small dataset. This is particularly useful in healthcare settings where labelled data is scarce but high-quality. *iii) Domain adaptation.* This method fine-tunes the model to adapt to the specific language and terminology used in healthcare and physiological data. It helps the model understand medical jargon and abbreviations, improving its performance on healthcare-related tasks. *v) Prompt engineering.* By carefully designing prompts or instructions, the model can be guided to better understand and process physiological data. This method relies on the creativity of the prompts to elicit the desired response from the model without extensive retraining. By employing these fine-tuning methods, pre-trained LLMs can be effectively adapted to handle the nuances and complexities of physiological data, thereby improving their performance in healthcare applications and contributing to better patient outcomes.

In an article published in *Nature*, there is ongoing discussion about the capabilities future LLMs for health should possess (Moor et al., 2023). Although numerous conceptual frameworks have been proposed, there remains a need for rigorous experimental validation across various tasks, robustness analysis, and a deeper understanding of why universal representations are effective. This opens up ample opportunities for future research.

Operating independently from the utilisation of LLMs, another promising avenue of research is the development of foundational approaches using large-scale, unlabelled physiological data signals for health applications. This strategy entails harnessing the vast quantities of raw, unstructured data produced by healthcare systems and wearable technologies. The goal is to create models capable of understanding and interpreting complex physiological signals without relying extensively on labelled data. Key methodologies and focal areas for this research include representation learning (Bengio et al., 2013) and self-supervised learning (Liu et al., 2021b; Krishnan et al., 2022). Representation learning aims to discover universal data representations that are applicable across various tasks, such as diagnosing different conditions or predicting health outcomes. By extracting generalisable features from extensive datasets, models can be adapted more easily to specific healthcare applications with minimal fine-tuning required. Self-supervised learning, a branch of unsupervised learning, allows a model to generate its own supervisory signals from the input data. In the context of physiological signals, this could involve tasks like predicting subsequent sequences in a time series or identifying anomalies in heartbeat patterns. Such capabilities enable models to derive meaningful representations of health-related data.

Recently, we have observed the emergence of several foundation models for physiological signals, primarily developed by major tech companies such as Google and Apple Inc. These companies have access to vast datasets and possess the necessary computational resources. For instance, Apple has utilised millions of wearable PPG and ECG samples to pre-train a model for representing physiological signals (Abbaspourazad et al., 2023). Similarly, Google has em-

ployed thousands of hours of respiratory sound recordings to develop a foundation model for respiratory audio analysis using constructive learning approaches (Baur et al., 2024). These models serve as potent feature extractors for a variety of downstream tasks.

Training a foundation model demands extensive datasets and computational resources. A practical approach to realising this is through decentralised learning. Building directly upon the work presented in this thesis, edge devices or hospitals possessing physiological data can collaborate to develop foundational models without the need to exchange their private data, utilising federated learning. This approach maximises data and resource utilisation.

Overall, developing a foundation model that can handle diverse physiological data and effectively address multiple health-related tasks holds tremendous potential. It calls for further research, experimentation, and analysis to achieve robust and reliable results that can revolutionise health diagnostics and applications.

This marks the conclusion of my thesis, but not the end of my research journey. The intersection of machine learning and healthcare is a promising and socially significant direction. I am committed to channelling my efforts towards making further contributions.

Bibliography

Salar Abbaspourazad, Oussama Elachqar, Andrew Miller, Saba Emrani, Udhyakumar Nallasamy, and Ian Shapiro. Large-scale training of foundation models for wearable biosignals. In *Proceedings of the 12th International Conference on Learning Representations (ICLR)*, pages 1–12, 2023.

Moloud Abdar, Farhad Pourpanah, Sadiq Hussain, Dana Rezazadegan, Li Liu, Mohammad Ghavamzadeh, Paul Fieguth, Xiaochun Cao, Abbas Khosravi, U Rajendra Acharya, et al. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information Fusion*, 1(1):1–11, 2021.

Sitara Afzal, Muazzam Maqsood, Faria Nazir, Umair Khan, Farhan Aadil, Khalid M Awan, Irfan Mehmood, and Oh-Young Song. A data augmentation-based framework to handle class imbalance problem for alzheimer's stage detection. *IEEE Access*, 7(1):115528–115539, 2019.

Rayo Akande, James Brimicombe, Martin R Cowie, Andrew Dymond, Hannah Clair Lindén, Gregory YH Lip, Jenny Lund, Jonathan Mant, Madhumitha Pandiaraja, Emma Svennberg, et al. Characterising RR intervals in atrial fibrillation detected through screening. pages 1–4, 2023.

Erick A Perez Alday, Annie Gu, Amit J Shah, Chad Robichaux, An-Kwok Ian Wong, Chengyu Liu, Feifei Liu, Ali Bahrami Rad, Andoni Elola, Salman Seyedi, et al. Classification of 12-lead ECGs: The physionet/computing in cardiology challenge 2020. *Physiological Measurement*, 41(12):1–10, 2020.

Naser Alfed, Fouad Khelifi, Ahmed Bouridane, and Huseyin Seker. Pigment network-based skin cancer detection. In *Proceedings of the 37th annual international conference of the IEEE engineering in medicine and biology society (EMBC)*, pages 7214–7217. IEEE, 2015.

Shalbbya Ali, Safdar Tanweer, Syed Khalid, and Naseem Rao. Mel frequency cepstral coefficient: A review. In *Proceedings of the 2nd International Conference on ICT for Digital, Smart, and Sustainable Development (ICIDSSD)*, 2021.

Gokhan Altan, Yakup Kutlu, and Novruz Allahverdi. Deep learning on computerized analysis of

chronic obstructive pulmonary disease. *IEEE Journal of Biomedical and Health Informatics*, 24(5):1344–1350, 2019.

Zachi I Attia, Peter A Noseworthy, Francisco Lopez-Jimenez, Samuel J Asirvatham, Abhishek J Deshmukh, Bernard J Gersh, Rickey E Carter, Xiaoxi Yao, Alejandro A Rabinstein, Brad J Erickson, et al. An artificial intelligence-enabled ECG algorithm for the identification of patients with atrial fibrillation during sinus rhythm: A retrospective analysis of outcome prediction. *The Lancet*, 394(10201):861–867, 2019.

Yuri Sousa Aurelio, Gustavo Matheus de Almeida, Cristiano Leite de Castro, and Antonio Padua Braga. Learning from imbalanced data sets with weighted cross-entropy function. *Neural Processing Letters*, 50(2):1937–1949, 2019.

Roberta Avanzato and Francesco Beritelli. Automatic ECG diagnosis using convolutional neural network. *Electronics*, 9(6):951–959, 2020.

Pierre Baldi and Peter J Sadowski. Understanding dropout. pages 2814–2822, 2013.

Catarina Barata, M Emre Celebi, and Jorge S Marques. A survey of feature extraction in dermoscopy image analysis of skin cancer. *IEEE Journal of Biomedical and Health Informatics*, 23(3):1096–1109, 2018.

Dalal Bardou, Kun Zhang, and Sayed Mohammad Ahmad. Lung sounds classification using convolutional neural networks. *Artificial Intelligence in Medicine*, 88:58–69, 2018.

Sebastien Baur, Zaid Nabulsi, Wei-Hung Weng, Jake Garrison, Louis Blankemeier, Sam Fishman, Christina Chen, Sujay Kakarmath, Minyoi Maimbolwa, Nsala Sanjase, et al. Hear–health acoustic representations. *arXiv preprint arXiv:2403.02522*, 2024.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8): 1798–1828, 2013.

Viktor Bengs, Eyke Hüllermeier, and Willem Waegeman. Pitfalls of epistemic uncertainty quantification through loss minimisation. In *Proceedings of the 36th International Conference on Neural Information Processing Systems (NeurIPS)*, 2022.

Umang Bhatt, Javier Antorán, Yunfeng Zhang, Q Vera Liao, Prasanna Sattigeri, Riccardo Fogliato, Gabrielle Melançon, Ranganath Krishnan, Jason Stanley, Omesh Tickoo, et al. Uncertainty as a form of transparency: Measuring, communicating, and using uncertainty. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (AIES)*, pages 401–413, 2021.

David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877, 2017.

Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural network. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, pages 1613–1622, 2015.

Abraham Bohadana, Gabriel Izbicki, and Steve S Kraman. Fundamentals of lung auscultation. *New England Journal of Medicine*, 370(8):744–751, 2014.

Erika Bondareva, Tong Xia, Jing Han, and Cecilia Mascolo. Towards uncertainty-aware murmur detection in heart sounds via tandem learning. In *Proceedings of the Computing in Cardiology (CinC)*, pages 1–4. IEEE, 2022.

Chloë Brown, Jagmohan Chauhan, Andreas Grammenos, Jing Han, Apinan Hasthanasombat, Dimitris Spathis, Tong Xia, Pietro Cicuta, and Cecilia Mascolo. Exploring automatic diagnosis of COVID-19 from crowdsourced respiratory sound data. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD)*, pages 3474–3484, 2020a.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems (NeurIPS)*, pages 1877–1901, 2020b.

Han Cai, Chuang Gan, Ligeng Zhu, and Song Han. Tinytl: Reduce memory, not parameters for efficient on-device learning. In *Proceedings of the 34th International Conference on Neural Information Processing Systems (NeurIPS)*, pages 11285–11297, 2020.

Yue Cao, Thomas Andrew Geddes, Jean Yee Hwa Yang, and Pengyi Yang. Ensemble deep learning in bioinformatics. *Nature Machine Intelligence*, 2(9):500–508, 2020.

C Carissimo, G Cerro, H Debelle, E Packer, AJ Yarnall, L Rochester, Lisa Alcock, L Ferrigno, A Marino, T Di Libero, et al. Enhancing remote monitoring and classification of motor state in parkinson's disease using wearable technology and machine learning. In *Proceedings of the IEEE International Symposium on Medical Measurements and Applications (MeMeA)*, pages 1–6. IEEE, 2023.

Ahmad Chaddad, Jihao Peng, Jian Xu, and Ahmed Bouridane. Survey of explainable ai techniques in healthcare. *Sensors*, 23(2):634–660, 2023.

Bertrand Charpentier, Daniel Zügner, and Stephan Günnemann. Posterior network: Uncertainty estimation without ood samples via density-based pseudo-counts. In *Proceedings of the 34th*

International Conference on Neural Information Processing Systems (NeurIPS), pages 1356–1367, 2020.

Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16 (1):321–357, 2002.

Daoyuan Chen, Dawei Gao, Yuexiang Xie, Xuchen Pan, Zitao Li, Yaliang Li, Bolin Ding, and Jingren Zhou. Fs-real: Towards real-world cross-device federated learning. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, pages 3829–3841, 2023a.

Haokun Chen, Ahmed Frikha, Denis Krompass, Jindong Gu, and Volker Tresp. Fraug: Tackling federated learning with non-iid features via representation augmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4849–4859, 2023b.

Tianqi Chen, Emily Fox, and Carlos Guestrin. Stochastic gradient hamiltonian monte carlo. In *Proceedings of the 31st International Conference on Machine Learning (ICML)*, pages 1683–1691. PMLR, 2014.

Yanping Chen, Yuan Hao, Thanawin Rakthanmanon, Jesin Zakaria, Bing Hu, and Eamonn Keogh. A general framework for never-ending learning from time series streams. *Data Mining and Knowledge Discovery*, 29:1622–1664, 2015.

Junghwan Cho, Kyewook Lee, Ellie Shin, Garry Choy, and Synho Do. How much data is needed to train a medical image deep learning system to achieve necessary high accuracy? *arXiv preprint arXiv:1511.06348*, 2015.

Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014.

Noel CF Codella, David Gutman, M Emre Celebi, Brian Helba, Michael A Marchetti, Stephen W Dusza, Aadi Kalloo, Konstantinos Liopyris, Nabin Mishra, Harald Kittler, et al. Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi). In *Proceedings of the 2018 IEEE 15th international symposium on biomedical imaging (ISBI)*, pages 168–172, 2018.

Katherine M Collins, Umang Bhatt, and Adrian Weller. Eliciting and learning with soft labels from every annotator. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing (HCOMP)*, pages 40–52, 2022.

Thomas M Cover. *Elements of information theory*. John Wiley & Sons, 1999.

Graham Crow, Rose Wiles, Sue Heath, and Vikki Charles. Research ethics and data quality: The implications of informed consent. *International Journal of Social Research Methodology*, 9 (2):83–95, 2006.

Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9268–9277, 2019.

Ting Dang, Jing Han, Tong Xia, Dimitris Spathis, Erika Bondareva, Chloë Siegele-Brown, Jagmohan Chauhan, Andreas Grammenos, Apinan Hasthanasombat, R Andres Floto, et al. Exploring longitudinal cough, breath, and voice data for COVID-19 progression prediction via sequential deep learning: Model development and validation. *Journal of Medical Internet Research*, 24(6):1–12, 2022.

Ting Dang, Jing Han, Tong Xia, Erika Bondareva, Chloë Siegele-Brown, Jagmohan Chauhan, Andreas Grammenos, Dimitris Spathis, Pietro Cicuta, and Cecilia Mascolo. Conditional neural ode processes for individual disease progression forecasting: A case study on COVID-19. *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, 2023.

Trung Dang, Om Thakkar, Swaroop Ramaswamy, Rajiv Mathews, Peter Chin, and Françoise Beaufays. Revealing and protecting labels in distributed training. In *Proceedings of the 35th International Conference on Neural Information Processing Systems (NeurIPS)*, volume 34, pages 1727–1738, 2021.

Ittai Dayan, Holger R Roth, Aoxiao Zhong, Ahmed Harouni, Amilcare Gentili, Anas Z Abidin, Andrew Liu, Anthony Beardsworth Costa, Bradford J Wood, Chien-Sung Tsai, et al. Federated learning for predicting clinical outcomes in patients with COVID-19. *Nature Medicine*, 27(10):1735–1743, 2021.

Li Deng, Dong Yu, et al. Deep learning: Methods and applications. *Foundations and Trends in Signal Processing*, 7(3):197–387, 2014.

Gauri Deshpande and Björn Schuller. An overview on audio, signal, speech, & language processing for COVID-19. *arXiv preprint arXiv:2005.08579*, 2020.

Sauptik Dhar, Junyao Guo, Jiayi Liu, Samarth Tripathi, Unmesh Kurup, and Mohak Shah. A survey of on-device machine learning: An algorithms and learning theory perspective. *ACM Transactions on Internet of Things*, 2(3):1–49, 2021.

Thomas J DiCiccio and Bradley Efron. Bootstrap confidence intervals. *Statistical Science*, 11 (3):189–228, 1996.

Thomas G Dietterich. Ensemble methods in machine learning. In *Proceedings of the International Workshop on Multiple Classifier Systems*, pages 1–15. Springer, 2000.

Alexey Dosovitskiy and Thomas Brox. Inverting visual representations with convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4829–4837, 2016.

Qi Dou, Tiffany Y So, Meirui Jiang, Quande Liu, Varut Vardhanabhuti, Georgios Kaissis, Zeju Li, Weixin Si, Heather HC Lee, Kevin Yu, et al. Federated deep learning for detecting COVID-19 lung abnormalities in CT: A privacy-preserving multinational validation study. *NPJ Digital Medicine*, 4(1):1–11, 2021.

Krishnamurthy Dvijotham, Jim Winkens, Melih Barsbey, Sumedh Ghaisas, Nick Pawlowski, Robert Stanforth, Patricia MacWilliams, Zahra Ahmed, Shekoofeh Azizi, Yoram Bachrach, et al. Enhancing the reliability and accuracy of ai-enabled diagnosis via complementarity-driven deferral to clinicians. *Nature Medicine*, 29(1):1814–1820, 2023.

Emadeldeen Ahmed Ibrahim Ahmed Eldele. *Towards robust and label-efficient time series representation learning*. PhD thesis, Nanyang Technological University, 2023.

Florian Eyben, Martin Wöllmer, and Björn Schuller. Opensmile: The munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM International Conference on Multimedia (MM)*, pages 1459–1462, 2010.

Ines Feki, Sourour Ammar, Yousri Kessentini, and Khan Muhammad. Federated learning for COVID-19 screening from chest X-Ray images. *Applied Soft Computing*, 106(1):1–12, 2021.

Meng Feng, Chieh-Chi Kao, Qingming Tang, Ming Sun, Viktor Rozgic, Spyros Matsoukas, and Chao Wang. Federated self-supervised learning for acoustic event classification. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 481–485, 2022a.

Tao Feng, Tong Xia, Xiaochen Fan, Huandong Wang, Zefang Zong, and Yong Li. Precise mobility intervention for epidemic control using unobservable information via deep reinforcement learning. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, pages 2882–2892, 2022b.

Tao Feng, Sirui Song, Tong Xia, and Yong Li. Contact tracing and epidemic intervention via deep reinforcement learning. *ACM Transactions on Knowledge Discovery from Data*, 17(3): 1–24, 2023.

Kenneth A Fleming, Susan Horton, Michael L Wilson, Rifat Atun, Kristen DeStigter, John Flanigan, Shahin Sayed, Pierrick Adam, Bertha Aguilar, Savvas Andronikou, et al. The lancet commission on diagnostics: transforming access to diagnostics. *The Lancet*, 398(10315): 1997–2050, 2021.

Mohammad Fraiwan, Luay Fraiwan, Basheer Khassawneh, and Ali Ibnian. A dataset of lung sounds recorded from the chest wall using an electronic stethoscope. *Data in Brief*, 35:10–20, 2021.

Siddhartha Gairola, Francis Tom, Nipun Kwatra, and Mohit Jain. Respirenet: A deep neural network for accurately detecting abnormal lung sounds in limited data setting. In *Proceedings of the 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 527–530, 2021.

Yarin Gal and Zoubin Ghahramani. Bayesian convolutional neural networks with bernoulli approximate variational inference. *arXiv preprint arXiv:1506.02158*, 2015.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of the 33rd International Conference on Machine Learning (ICML)*, pages 1050–1059, 2016a.

Yarin Gal and Zoubin Ghahramani. A theoretically grounded application of dropout in recurrent neural networks. pages 1019–1027, 2016b.

Mudasir A Ganaie, Minghui Hu, Ashwani Kumar Malik, Muhammad Tanveer, and Ponnuthurai N Suganthan. Ensemble deep learning: A review. *Engineering Applications of Artificial Intelligence*, 115(1):105–151, 2022.

Yan Gao, Titouan Parcollet, Salah Zaiem, Javier Fernandez-Marques, Pedro PB de Gusmao, Daniel J Beutel, and Nicholas D Lane. End-to-end speech recognition from federated acoustic models. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7227–7231, 2022.

Jakob Gawlikowski, Cedrique Rovile Njieutcheu Tassi, Mohsin Ali, Jongseok Lee, Matthias Humt, Jianxiang Feng, Anna Kruspe, Rudolph Triebel, Peter Jung, Ribana Roscher, et al. A survey of uncertainty in deep neural networks. *arXiv preprint arXiv:2107.03342*, 2021.

Abhirup Ghosh and Tong Xia. Mobility-based individual poi recommendation to control the COVID-19 spread. In *Proceedings of the IEEE International Conference on Big Data (Big Data)*, pages 4356–4364. IEEE, 2021.

Biraja Ghoshal and Allan Tucker. Estimating uncertainty and interpretability in deep learning for coronavirus (COVID-19) detection. *arXiv preprint arXiv:2003.10769*, 2020.

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

Alex Graves. Practical variational inference for neural networks. In *Proceedings of the 15th International Conference on Neural Information Processing Systems (NeurIPS)*, pages 2348–2356. Citeseer, 2011.

Alex Graves and Alex Graves. Long short-term memory. *Supervised Sequence Labelling with Recurrent Neural Networks*, 1(1):37–45, 2012.

Flavia Grignaffini, Francesco Barbuto, Lorenzo Piazzo, Maurizio Troiano, Patrizio Simeoni, Fabio Mangini, Giovanni Pellacani, Carmen Cantisani, and Fabrizio Frezza. Machine learning approaches for skin cancer classification from dermoscopic images: A systematic review. *Algorithms*, 15(11):438, 2022.

Jiuxiang Gu, Zhenhua Wang, Jason Kuen, Lianyang Ma, Amir Shahroudy, Bing Shuai, Ting Liu, Xingxing Wang, Gang Wang, Jianfei Cai, et al. Recent advances in convolutional neural networks. *Pattern Recognition*, 77(1):354–377, 2018.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *Proceedings of the 29th International Conference on Machine Learning (ICML*, pages 1321–1330. PMLR, 2017.

Yongxin Guo, Xiaoying Tang, and Tao Lin. Fedbr: Improving federated learning on heterogeneous data via local learning bias reduction. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, pages 12034–12054. PMLR, 2023.

Arjun K Gupta and Saralees Nadarajah. *Handbook of beta distribution and its applications*. CRC Press, 2004.

Stefan Hadjitodorov and Petar Mitev. A computer system for acoustic analysis of pathological voices and laryngeal diseases screening. *Medical Engineering & Physics*, 24(6):419–429, 2002.

Jing Han, Zixing Zhang, Maximilian Schmitt, Maja Pantic, and Björn Schuller. From hard to soft: Towards more human-like emotion recognition by modelling the perception uncertainty. In *Proceedings of the 25th ACM International Conference on Multimedia (MM)*, pages 890–897, 2017.

Jing Han, Kun Qian, Meishu Song, Zijiang Yang, Zhao Ren, Shuo Liu, Juan Liu, Huaiyuan Zheng, Wei Ji, Tomoya Koike, et al. An early study on intelligent analysis of speech un-

der COVID-19: Severity, sleep quality, fatigue, and anxiety. In *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 4946–4950, 2020.

Jing Han, Chloë Brown, Jagmohan Chauhan, Andreas Grammenos, Apinan Hasthanasombat, Dimitris Spathis, Tong Xia, Pietro Cicuta, and Cecilia Mascolo. Exploring automatic COVID-19 diagnosis via voice and symptoms from crowdsourced data. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8328–8332. IEEE, 2021a.

Jing Han, Tong Xia, Dimitris Spathis, Erika Bondareva, Chloë Brown, Jagmohan Chauhan, Ting Dang, Andreas Grammenos, Apinan Hasthanasombat, Andres Floto, et al. Sounds of COVID-19: Exploring realistic performance of audio-based digital testing. *NPJ Digital Medicine*, 5 (1):16–26, 2022.

Jing Han, Marco Montagna, Andreas Grammenos, Tong Xia, Erika Bondareva, Chloë Siegele-Brown, Jagmohan Chauhan, Ting Dang, Dimitris Spathis, R Andres Floto, et al. Evaluating listening performance for COVID-19 detection by clinicians and machine learning: comparative study. *Journal of Medical Internet Research*, 25:1–10, 2023a.

Xudong Han, Timothy Baldwin, and Trevor Cohn. Diverse adversaries for mitigating bias in training. *arXiv preprint arXiv:2101.10001*, 2021b.

Zhenyu Han, Tong Xia, Yanxin Xi, and Yong Li. Healthy cities, a comprehensive dataset for environmental determinants of health in england cities. *Scientific Data*, 10(1):165–175, 2023b.

Ibrahim R Hanna and Mark E Silverman. A history of cardiac auscultation and some of its contributors. *The American journal of cardiology*, 90(3):259–267, 2002.

Tian Hao, Guoliang Xing, and Gang Zhou. isleep: Unobtrusive sleep quality monitoring using smartphones. In *Proceedings of the 11th ACM Conference on Embedded Networked Sensor Systems (SenSys)*, pages 1–14, 2013.

Fredric J Harris. On the use of windows for harmonic analysis with the discrete fourier transform. *Proceedings of the IEEE*, 66(1):51–83, 1978.

Trevor Hastie, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman. *The elements of statistical learning: Data mining, inference, and prediction*, volume 2. Springer, 2009.

Faezeh Nejati Hatamian, Nishant Ravikumar, Sulaiman Vesal, Felix P Kemeth, Matthias Struck, and Andreas Maier. The effect of data augmentation on classification of atrial fibrillation in short single-lead ECG signals using deep neural networks. In *Proceedings of the IEEE*

International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 1264–1268. IEEE, 2020.

Haibo He, Yang Bai, Edwardo A Garcia, and Shutao Li. Adasyn: Adaptive synthetic sampling approach for imbalanced learning. In *Proceedings of the IEEE International Joint Conference on Neural Networks (IJCNN)*, pages 1322–1328. IEEE, 2008.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

Robert Hecht-Nielsen. Theory of the backpropagation neural network. In *Neural Networks for Perception*, volume 1, pages 65–93. Elsevier, 1992.

Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al. CNN architectures for large-scale audio classification. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 131–135, 2017.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Jiri Hron, Alexander G de G Matthews, and Zoubin Ghahramani. Variational gaussian dropout is not bayesian. *arXiv preprint arXiv:1711.02989*, 2017.

Chao Huang, Jianwei Huang, and Xin Liu. Cross-silo federated learning: Challenges and opportunities. *arXiv preprint arXiv:2206.12949*, 2022a.

Yinghui Huang, Sijun Meng, Yi Zhang, et al. The respiratory sound features of COVID-19 patients fill gaps between clinical data and screening methods. *medRxiv preprint: 2020.04.07.20051060*, 2020.

Yingsong Huang, Bing Bai, Shengwei Zhao, Kun Bai, and Fei Wang. Uncertainty-aware learning against label noise on imbalanced datasets. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pages 11147–11155, 2022b.

Tove Hygrell, Fredrik Viberg, Erik Dahlberg, Peter H Charlton, Katrin Kemp Gudmundsdottir, Jonathan Mant, Josef Lindman Hörnlund, and Emma Svennberg. An artificial intelligence-based model for prediction of atrial fibrillation from single-lead sinus rhythm electrocardiograms facilitating screening. *Europace*, 25(4):1332–1338, 2023.

Ali Imran, Iryna Posokhova, Haneya N Qureshi, Usama Masood, Muhammad Sajid Riaz, Kamran Ali, Charles N John, MD Iftikhar Hussain, and Muhammad Nabeel. AI4 COVID-19: AI

enabled preliminary diagnosis for COVID-19 from cough samples via an app. *Informatics in Medicine Unlocked*, 20(1):100–1110, 2020.

Abinaya Inbamani, A. Siva Sakthi, R.R. Rubia Gandhi, M. Preethi, R. Rajalakshmi, Veerapandi Veerasamy, and Thirumeni Mariammal. *Chapter 2 - Role of IoT and semantics in e-Health*. Academic Press, 2022.

Imran Iqbal, Muhammad Younus, Khuram Walayat, Mohib Ullah Kakar, and Jinwen Ma. Automated multi-class classification of skin lesions through deep convolutional neural network with dermoscopic images. *Computerized Medical Imaging and Graphics*, 88:101843, 2021.

Hitoshi Iyatomi, M Emre Celebi, Gerald Schaefer, and Masaru Tanaka. Automated color calibration method for dermoscopy images. *Computerized Medical Imaging and Graphics*, 35 (2):89–98, 2011.

Péter Jacsó. Google scholar: the pros and the cons. *Online Information Review*, 1(1):1–10, 2005.

Shweta H Jambukia, Vipul K Dabhi, and Harshadkumar B Prajapati. Classification of ECG signals using machine learning techniques: A survey. In *Proceedings of the International Conference on Advances in Computer Engineering and Applications (ICACEA)*, pages 714–721. IEEE, 2015.

Rabia Javed, Mohd Shafry Mohd Rahim, Tanzila Saba, and Amjad Rehman. A comparative study of features selection for skin lesion detection from dermoscopic images. *Network Modeling Analysis in Health Informatics and Bioinformatics*, 9(1):4, 2020.

Eunjeong Jeong, Seungeun Oh, Hyesung Kim, Jihong Park, Mehdi Bennis, and Seong-Lyun Kim. Communication-efficient on-device machine learning: Federated distillation and augmentation under non-iid private data. *arXiv preprint arXiv:1811.11479*, 2018.

Justin M Johnson and Taghi M Khoshgoftaar. Survey on deep learning with class imbalance. *Journal of Big Data*, 6(1):1–54, 2019.

Michael I Jordan and Tom M Mitchell. Machine learning: Trends, perspectives, and prospects. *Science*, 349(6245):255–260, 2015.

Cheng Ju, Aurélien Bibaut, and Mark van der Laan. The relative performance of ensemble methods with deep convolutional neural networks for image classification. *Journal of Applied Statistics*, 45(15):2800–2818, 2018.

Dae Y Kang, Pamela N DeYoung, Justin Tantiongloc, Todd P Coleman, and Robert L Owens. Statistical uncertainty quantification to augment clinical decision support: A first implementation in sleep medicine. *NPJ Digital Medicine*, 4(1):142, 2021.

Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha Suresh. Scaffold: Stochastic controlled averaging for federated learning. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, pages 5132–5143. PMLR, 2020.

Sai Praneeth Karimireddy, Martin Jaggi, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian U Stich, and Ananda Theertha Suresh. Breaking the centralized barrier for cross-device federated learning. In *Proceedings of the 35th International Conference on Neural Information Processing Systems (NeurIPS)*, pages 28663–28676, 2021.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Harold Kittler, H Pehamberger, K Wolff, and MJTIO Binder. Diagnostic accuracy of dermoscopy. *The Lancet Oncology*, 3(3):159–165, 2002.

Paul Kligfield, Leonard S Gettes, James J Bailey, Rory Childers, Barbara J Deal, E William Hancock, Gerard Van Herpen, Jan A Kors, Peter Macfarlane, David M Mirvis, et al. Recommendations for the standardization and interpretation of the electrocardiogram. *Circulation*, 115(10):1306–1324, 2007.

Bee Hock David Koh. *Signal processing and analytics of multimodal biosignals*. PhD thesis, Newcastle University, 2019.

Anna-Kathrin Kopetzki, Bertrand Charpentier, Daniel Zügner, Sandhya Giri, and Stephan Günnemann. Evaluating robustness of predictive uncertainty estimation: Are dirichlet-based models reliable? In *Proceedings of the 33th International Conference on Machine Learning (ICML)*, pages 5707–5718, 2021.

Frauke Kreuter, Georg-Christoph Haas, Florian Keusch, Sebastian Bähr, and Mark Trappmann. Collecting survey and smartphone sensor data with an app: Opportunities and challenges around privacy and informed consent. *Social Science Computer Review*, 38(5):533–549, 2020.

Rayan Krishnan, Pranav Rajpurkar, and Eric J Topol. Self-supervised learning in medicine and healthcare. *Nature Biomedical Engineering*, 6(12):1346–1352, 2022.

Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. Cifar-10 and cifar-100 datasets. *URL: https://www.cs.toronto.edu/kriz/cifar.html*, 2009.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NeurIPS)*, pages 6405–6416, 2017.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

Gihun Lee, Minchan Jeong, Yongjin Shin, Sangmin Bae, and Se-Young Yun. Preservation of the global knowledge by not-true distillation in federated learning. In *Proceedings of the 36th International Conference on Neural Information Processing Systems (NeurIPS)*, 2022.

Kimin Lee, Honglak Lee, Kibok Lee, and Jinwoo Shin. Training confidence-calibrated classifiers for detecting out-of-distribution samples. In *Proceedings of the International Conference on Learning Representations (ICLR)*, pages 1–10, 2018.

Christian Leibig, Vaneeda Allken, Murat Seçkin Ayhan, Philipp Berens, and Siegfried Wahl. Leveraging uncertainty information from deep neural networks for disease detection. *Scientific Reports*, 7(1):1–14, 2017.

Andreanne Lemay, Katharina Hoebel, Christopher P Bridge, Brian Befano, Silvia De Sanjosé, Didem Egemen, Ana Cecilia Rodriguez, Mark Schiffman, John Peter Campbell, and Jayashree Kalpathy-Cramer. Improving the repeatability of deep learning models with monte carlo dropout. *npj Digital Medicine*, 5(1):1–11, 2022.

Daniel Levy, Yair Carmon, John C Duchi, and Aaron Sidford. Large-scale methods for distributionally robust optimization. In *Proceedings of the 34th International Conference on Neural Information Processing Systems (NeurIPS)*, pages 8847–8860, 2020.

Hao Li, Yang Nan, Javier Del Ser, and Guang Yang. Region-based evidential deep learning to quantify uncertainty and improve robustness of brain tumor segmentation. *Neural Computing and Applications*, 35(30):22071–22085, 2023.

Li Li, Yuxi Fan, Mike Tse, and Kuo-Yi Lin. A review of applications in federated learning. *Computers & Industrial Engineering*, 149:106–117, 2020a.

Tian Li, Anit Kumar Sahu, Ameet Talwalkar, and Virginia Smith. Federated learning: Challenges, methods, and future directions. *IEEE Signal Processing Magazine*, 37(3):50–60, 2020b.

Tong Li, Tong Xia, Huandong Wang, Zhen Tu, Sasu Tarkoma, Zhu Han, and Pan Hui. Smartphone app usage analysis: Datasets, methods, and applications. *IEEE Communications Surveys & Tutorials*, 24(2):937–966, 2022.

Xiang Li, Kaixuan Huang, Wenhao Yang, Shusen Wang, and Zhihua Zhang. On the convergence of fedavg on non-iid data. *arXiv preprint arXiv:1907.02189*, 2019.

Xiaoxiao Li, Meirui JIANG, Xiaofei Zhang, Michael Kamp, and Qi Dou. Fedbn: Federated learning on non-iid features via local batch normalization. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2020c.

Xin-Chun Li and De-Chuan Zhan. Fedrs: Federated learning with restricted softmax for label distribution non-iid data. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining (KDD)*, pages 995–1005, 2021.

Yuchong Li and Qinghui Liu. A comprehensive review study of cyber-attacks and cyber security: Emerging trends and recent developments. *Energy Reports*, 7(1):8176–8186, 2021.

Wei Yang Bryan Lim, Nguyen Cong Luong, Dinh Thai Hoang, Yutao Jiao, Ying-Chang Liang, Qiang Yang, Dusit Niyato, and Chunyan Miao. Federated learning in mobile edge networks: A comprehensive survey. *IEEE Communications Surveys & Tutorials*, 22(3):2031–2063, 2020.

Daoqin Lin, Yuchun Guo, Huan Sun, and Yishuai Chen. Fedcluster: A federated learning framework for cross-device private ECG classification. In *Proceedings of the IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, pages 1–6, 2022a.

Ji Lin, Ligeng Zhu, Wei-Ming Chen, Wei-Chen Wang, Chuang Gan, and Song Han. On-device training under 256kb memory. In *Proceedings of the 36th International Conference on Neural Information Processing Systems (NeurIPS)*, pages 22941–22954, 2022b.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision (CVPR)*, pages 2980–2988, 2017.

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

Gaoyang Liu, Chen Wang, Xiaoqiang Ma, and Yang Yang. Keep your data locally: Federated-learning-based data privacy preservation in edge computing. *IEEE Network*, 35(2):60–66, 2021a.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35, 2023a.

Shunjian Liu, Xinxin Feng, and Haifeng Zheng. Overcoming forgetting in local adaptation of

federated learning model. In *Proceedings of the Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining (PAKDD)*, pages 613–625. Springer, 2022.

Xiao Liu, Fanjin Zhang, Zhenyu Hou, Li Mian, Zhaoyu Wang, Jing Zhang, and Jie Tang. Self-supervised learning: Generative or contrastive. *IEEE Transactions on Knowledge and Data Engineering*, 35(1):857–876, 2021b.

Xin Liu, Daniel McDuff, Geza Kovacs, Isaac Galatzer-Levy, Jacob Sunshine, Jiening Zhan, Ming-Zher Poh, Shun Liao, Paolo Di Achille, and Shwetak Patel. Large language models are few-shot health learners. *arXiv preprint arXiv:2305.15525*, 2023b.

Yi Liu, Chao Yang, Zengliang Gao, and Yuan Yao. Ensemble deep kernel learning with application to quality prediction in industrial polymerization processes. *Chemometrics and Intelligent Laboratory Systems*, 174:15–21, 2018.

Adria Romero Lopez, Xavier Giro-i Nieto, Jack Burdick, and Oge Marques. Skin lesion classification from dermoscopic images using deep learning techniques. In *Proceedings of the 13th International Conference on Biomedical Engineering (BioMed)*, pages 49–54. IEEE, 2017.

Romain Lopez, Jeffrey Regier, Michael I Jordan, and Nir Yosef. Information constraints on auto-encoding variational bayes. pages 31–40, 2018.

Christos Louizos and Max Welling. Multiplicative normalizing flows for variational bayesian neural networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, pages 2218–2227. PMLR, 2017.

Mi Luo, Fei Chen, Dapeng Hu, Yifan Zhang, Jian Liang, and Jiashi Feng. No fear of heterogeneity: Classifier calibration for federated learning with non-iid data. pages 5972–5984, 2021.

Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. Biogpt: generative pre-trained transformer for biomedical text generation and mining. *Briefings in Bioinformatics*, 23(6):1–10, 2022.

Xiaodong Ma, Jia Zhu, Zhihao Lin, Shanxuan Chen, and Yangjie Qin. A state-of-the-art survey on solving non-iid data in federated learning. *Future Generation Computer Systems*, 135(1): 244–258, 2022.

Andrey Malinin and Mark Gales. Predictive uncertainty estimation via prior networks. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems (NeurIPS)*, pages 6405–6416, 2018.

Tanis Mar, Sebastian Zaunseder, Juan Pablo Martínez, Mariano Llamedo, and Rüdiger Poll.

Optimization of ECG classification by means of feature selection. *IEEE Transactions on Biomedical Engineering*, 58(8):2168–2177, 2011.

Roman C Maron, Justin G Schlager, Sarah Haggenmüller, Christof von Kalle, Jochen S Utikal, Friedegund Meier, Frank F Gellrich, Sarah Hobelsberger, Axel Hauschild, Lars French, et al. A benchmark for neural network robustness in skin cancer classification. *European Journal of Cancer*, 155(1):191–199, 2021.

Cecilia Mascolo. Mobile health diagnostics through audio signals. In *Proceedings of Deep Learning for Wellbeing Applications Leveraging Mobile Devices and Edge Computing (HEALTHDL)*, pages 16–16, 2020.

Simon C Mathews, Michael J McShea, Casey L Hanley, Alan Ravitz, Alain B Labrique, and Adam B Cohen. Digital health: A path to validation. *NPJ Digital Medicine*, 2(1):38–40, 2019.

Maciej A Mazurowski, Piotr A Habas, Jacek M Zurada, Joseph Y Lo, Jay A Baker, and Georgia D Tourassi. Training neural network classifiers for medical decision making: The effects of imbalanced datasets on classification performance. *Neural Networks*, 21(2):427–436, 2008.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 1273–1282, 2017.

Larry R Medsker and LC Jain. Recurrent neural networks. *Design and Applications*, 5(1): 64–67, 2001.

Agnieszka Mikołajczyk, Arkadiusz Kwasigroch, and Michał Grochowski. Intelligent system supporting diagnosis of malignant melanoma. In *Proceedings of the Trends in Advanced Intelligent Control, Optimization and Automation (KKA)*, pages 828–837. Springer, 2017.

Michael Moor, Oishi Banerjee, Zahra Shakeri Hossein Abad, Harlan M Krumholz, Jure Leskovec, Eric J Topol, and Pranav Rajpurkar. Foundation models for generalist medical artificial intelligence. *Nature*, 616(7956):259–265, 2023.

Himadri Mukherjee, Priyanka Sreerama, Ankita Dhar, Sk Md Obaidullah, Kaushik Roy, Mufti Mahmud, and KC Santosh. Automatic lung health screening using respiratory sounds. *Journal of Medical Systems*, 45(2):1–9, 2021.

Kevin P Murphy. *Machine learning: A probabilistic perspective*. MIT press, 2012.

Mohsen Nabian, Athena Nouhi, Yu Yin, and Sarah Ostadabbas. A biosignal-specific processing

tool for machine learning and pattern recognition. In *Proceedings of the IEEE Healthcare Innovations and Point of Care Technologies (HI-POCT)*, pages 76–80. IEEE, 2017.

Radford M Neal. *Bayesian learning for neural networks*, volume 118. Springer Science & Business Media, 2012.

Dinh C Nguyen, Ming Ding, Pubudu N Pathirana, Aruna Seneviratne, Jun Li, and H Vincent Poor. Federated learning for internet of things: A comprehensive survey. *IEEE Communications Surveys & Tutorials*, 23(3):1622–1658, 2021a.

Duc-Khanh Nguyen, Chung-Hsien Lan, and Chien-Lung Chan. Deep ensemble learning approaches in healthcare to enhance the prediction and diagnosing performance: The workflows, deployments, and surveys on the statistical, image-based, and sequential datasets. *International Journal of Environmental Research and Public Health*, 18(20):1–12, 2021b.

Truc Nguyen and Franz Pernkopf. Lung sound classification using snapshot ensemble of convolutional neural networks. In *Proceedings of the 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 760–763. IEEE, 2020.

Alexandru Niculescu-Mizil and Rich Caruana. Predicting good probabilities with supervised learning. In *Proceedings of the 22nd International Conference on Machine Learning (ICML)*, pages 625–632, 2005.

Sebastian W Ober, Carl E Rasmussen, and Mark van der Wilk. The promises and pitfalls of deep kernel learning. In *Proceedings of the Uncertainty in Artificial Intelligence (UAI)*, pages 1206–1216. PMLR, 2021.

Robert Thomas Olszewski. *Generalized feature extraction for structural pattern recognition in time-series data*. Carnegie Mellon University, 2001.

Alan V Oppenheim. *Discrete-time signal processing*. Pearson Education India, 1999.

Christina Orphanidou. A review of big data applications of physiological signal data. *Biophysical reviews*, 11(1):83–87, 2019.

Christina Orphanidou, Timothy Bonnici, Peter Charlton, David Clifton, David Vallance, and Lionel Tarassenko. Signal-quality indices for the electrocardiogram and photoplethysmogram: Derivation and applications to wireless monitoring. *IEEE Journal of Biomedical and Health Informatics*, 19(3):832–8, 2015. ISSN 2168-2208.

Ian Osband. Risk versus uncertainty in deep learning: Bayes, bootstrap and the dangers of dropout. In *Proceedings of the NIPS workshop on Bayesian Deep Learning*, pages 192–210, 2016.

Keiron O'Shea and Ryan Nash. An introduction to convolutional neural networks. *arXiv preprint arXiv:1511.08458*, 2015.

Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, David Sculley, Sebastian Nowozin, Joshua Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems (NeurIPS)*, pages 1–13, 2019.

Şaban Öztürk and Tolga Çukur. Deep clustering via center-oriented margin free-triplet loss for skin lesion detection in highly imbalanced datasets. *IEEE Journal of Biomedical and Health Informatics*, 26(9):4679–4690, 2022.

Andre GC Pacheco, Chandramouli S Sastry, Thomas Trappenberg, Sageev Oore, and Renato A Krohling. On out-of-distribution detection algorithms with deep neural skin cancer classifiers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 732–733, 2020.

Chunjong Park, Anas Awadalla, Tadayoshi Kohno, and Shwetak Patel. Reliable and trustworthy machine learning for health using dataset shift detection. In *Proceedings of the 35th International Conference on Neural Information Processing Systems (NeurIPS)*, pages 1–13, 2021.

Annette Plüddemann, Carl Heneghan, Matthew Thompson, Jane Wolstenholme, and Christopher P Price. Dermoscopy for the diagnosis of melanoma: Primary care diagnostic technology update. *British Journal of General Practice*, 61(587):416–417, 2011.

Victor Pomponiu, Hossein Nejati, and N-M Cheung. Deepmole: Deep neural networks for skin mole lesion classification. In *Proceedings of the IEEE international conference on image processing (ICIP)*, pages 2623–2627. IEEE, 2016.

Janis Postels, Mattia Segu, Tao Sun, Luc Van Gool, Fisher Yu, and Federico Tombari. On the practicality of deterministic epistemic uncertainty. pages 1–10, 2022.

Renard Xaviero Adhi Pramono, Syed Anas Imtiaz, and Esther Rodriguez-Villegas. A cough-based algorithm for automatic diagnosis of pertussis. *PloS One*, 11(9):1–10, 2016.

Adnan Qayyum, Kashif Ahmad, Muhammad Ahtazaz Ahsan, Ala Al-Fuqaha, and Junaid Qadir. Collaborative federated learning for healthcare: Multi-modal COVID-19 diagnosis at the edge. *IEEE Open Journal of the Computer Society*, 1(1):172–184, 2022.

Lorena Qendro, Alexander Campbell, Pietro Lio, and Cecilia Mascolo. Early exit ensembles for uncertainty quantification. In *Proceedings of the Machine Learning for Health (ML4H)*, pages 181–195, 2021a.

Lorena Qendro, Jagmohan Chauhan, Alberto Gil CP Ramos, and Cecilia Mascolo. The benefit

of the doubt: Uncertainty aware sensing for edge computing platforms. In *Proceedings of the IEEE/ACM Symposium on Edge Computing (SEC)*, pages 214–227. IEEE, 2021b.

Maithra Raghu, Katy Blumer, Rory Sayres, Ziad Obermeyer, Bobby Kleinberg, Sendhil Mullainathan, and Jon Kleinberg. Direct uncertainty prediction for medical second opinions. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, pages 5281–5290. PMLR, 2019.

M Mostafizur Rahman and Darryl N Davis. Addressing the class imbalance problem in medical datasets. *International Journal of Machine Learning and Computing*, 3(2):224–234, 2013.

Jean-Francois Rajotte, Sumit Mukherjee, Caleb Robinson, Anthony Ortiz, Christopher West, Juan M Lavista Ferres, and Raymond T Ng. Reducing bias and increasing utility by federated generative modeling of medical images using a centralized adversary. In *Proceedings of the Conference on Information Technology for Social Good (GoodIT)*, pages 79–84, 2021.

Pranav Rajpurkar, Emma Chen, Oishi Banerjee, and Eric J Topol. Ai in health and medicine. *Nature Medicine*, 28(1):31–38, 2022.

Khalid Raza. Improving the prediction accuracy of heart disease with ensemble learning and majority voting rule. *U-Healthcare Monitoring Systems*, 1(2):179–196, 2019.

Sandeep Reddy, Sonia Allan, Simon Coghlan, and Paul Cooper. A governance model for the application of ai in health care. *Journal of the American Medical Informatics Association*, 27 (3):491–497, 2020.

Haoran Ren, Arnab Neelim Mazumder, Hasib-Al Rashid, Vandana Chandrareddy, Aidin Shiri, Nitheesh Kumar Manjunath, and Tinoosh Mohsenin. End-to-end scalable and low power multi-modal cnn for respiratory-related symptoms detection. In *Proceedings of the 33rd International System-on-Chip Conference (SOCC)*, pages 102–107. IEEE, 2020.

Nicola Rieke, Jonny Hancox, Wenqi Li, Fausto Milletari, Holger R Roth, Shadi Albarqouni, Spyridon Bakas, Mathieu N Galtier, Bennett A Landman, Klaus Maier-Hein, et al. The future of digital health with federated learning. *NPJ digital medicine*, 3(1):119–129, 2020.

Beanbonyka Rim, Nak-Jun Sung, Sedong Min, and Min Hong. Deep learning in physiological signal data: A survey. *Sensors*, 20(4):969–976, 2020.

Bruno M Rocha, Dimitris Filos, Luís Mendes, Gorkem Serbes, Sezer Ulukaya, Yasemin P Kahya, Nikša Jakovljevic, Tatjana L Turukalo, Ioannis M Vogiatzis, Eleni Perantoni, et al. An open access database for the evaluation of respiratory sound classification algorithms. *Physiological Measurement*, 40(3):1–10, 2019.

Johanna Rock, Tiago Azevedo, René de Jong, Daniel Ruiz-Muñoz, and Partha Maji. On efficient uncertainty estimation for resource-constrained mobile applications. *arXiv preprint arXiv:2111.09838*, 2021.

Frank Rosenblatt. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, 65(6):386–388, 1958.

Reuven Y Rubinstein and Dirk P Kroese. *The cross-entropy method: A unified approach to combinatorial optimization, Monte-Carlo simulation, and machine learning*, volume 133. Springer, 2004.

Aurora Saez, Carmen Serrano, and Begona Acha. Model-based classification methods of global patterns in dermoscopic images. *IEEE Transactions on Medical Imaging*, 33(5):1137–1147, 2014.

Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*, 2019.

Manisha Saini and Seba Susan. Data augmentation of minority class with transfer learning for classification of imbalanced breast cancer dataset using inception-v3. In *Proceedings of the Iberian Conference on Pattern Recognition and Image Analysis (PRIA)*, pages 409–420. Springer, 2019.

Elena Sajno, Sabrina Bartolotta, Cosimo Tuena, Pietro Cipresso, Elisa Pedroli, and Giuseppe Riva. Machine learning in biosignals processing for mental health: A narrative review. *Frontiers in Psychology*, 13(1):1–10, 2023.

Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4510–4520, 2018.

Deepti Saraswat, Pronaya Bhattacharya, Ashwin Verma, Vivek Kumar Prasad, Sudeep Tanwar, Gulshan Sharma, Pitshou N Bokoro, and Ravi Sharma. Explainable ai for healthcare 5.0: Opportunities and challenges. *IEEE Access*, 6(1):1–10, 2022.

Hassan Sarmadi, Alireza Entezami, Behzad Saeedi Razavi, and Ka-Veng Yuen. Ensemble learning-based structural health monitoring by mahalanobis distance metrics. *Structural Control and Health Monitoring*, 28(2):1–10, 2021.

Felix Sattler, Klaus-Robert Müller, Thomas Wiegand, and Wojciech Samek. On the byzantine robustness of clustered federated learning. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8861–8865, 2020.

Björn Schuller. Chain of audio processing. In *Intelligent Audio Analysis*, pages 17–22. Springer, 2013.

Sarah Schumacher, Winfried Rief, Elmar Brähler, Alexandra Martin, Heide Glaesmer, and Ricarda Mewes. Disagreement in doctor's and patient's rating about medically unexplained symptoms and health care use. *International Journal of Behavioral Medicine*, 20(1):30–37, 2013.

Murat Sensoy, Lance Kaplan, and Melih Kandemir. Evidential deep learning to quantify classification uncertainty. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems (NeurIPS)*, pages 6405–6416, 2018.

Roneel V Sharan. Cough sound detection from raw waveform using sincnet and bidirectional gru. *Biomedical Signal Processing and Control*, 82(1):1–10, 2023.

Mayuri S Shelke, Prashant R Deshmukh, and Vijaya K Shandilya. A review on imbalanced data handling using undersampling and oversampling technique. *International Journal of Recent Trends in Engineering and Research*, 3(4):444–449, 2017.

Maohao Shen, Yuheng Bu, Prasanna Sattigeri, Soumya Ghosh, Subhro Das, and Gregory Wornell. Post-hoc uncertainty learning using a dirichlet meta-model. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 37, pages 9772–9781, 2023.

Zebang Shen, Juan Cervino, Hamed Hassani, and Alejandro Ribeiro. An agnostic approach to federated learning with class imbalance. In *Proceedings of the International Conference on Learning Representations (ICLR)*, pages 1–12, 2021.

Lukui Shi, Kang Du, Chaozong Zhang, Hongqi Ma, and Wenjie Yan. Lung sound recognition algorithm based on vggish-bigru. *IEEE Access*, 7(1):139438–139449, 2019.

Neta Shoham, Tomer Avidor, Aviv Keren, Nadav Israel, Daniel Benditkis, Liron Mor-Yosef, and Itai Zeitak. Overcoming forgetting in federated learning on non-iid data. *arXiv preprint arXiv:1910.07796*, 2019.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Kusha Sridhar, Wei-Cheng Lin, and Carlos Busso. Generative approach using soft-labels to learn uncertainty in predicting emotional attributes. In *Proceedings of the International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 1–8. IEEE, 2021.

Arpan Srivastava, Sonakshi Jain, Ryan Miranda, Shruti Patil, Sharnil Pandya, and Ketan Kotecha. Deep learning based respiratory sound analysis for detection of chronic obstructive pulmonary disease. *PeerJ Computer Science*, 7(1), 2021.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

Steven R Steinhubl, Evan D Muse, and Eric J Topol. The emerging field of mobile health. *Science Translational Medicine*, 7(283):283–289, 2015.

Fei Sun, Jun Liu, Jian Wu, Changhua Pei, Xiao Lin, Wenwu Ou, and Peng Jiang. Bert4rec: Sequential recommendation with bidirectional encoder representations from transformer. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management (CIKM)*, pages 1441–1450, 2019.

Muhammad Shehram Shah Syed, Zafi Sherhan Syed, Margaret Lech, and Elena Pirogova. Automated screening for alzheimer's dementia through spontaneous speech. In *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 2222–6, 2020.

Lidia Talavera-Martinez, Pedro Bibiloni, and Manuel González-Hidalgo. Computational texture features of dermoscopic images and their link to the descriptive terminology: A survey. *Computer Methods and Programs in Biomedicine*, 182(1):1–10, 2019.

Srikanth Tammina. Transfer learning using vgg-16 with deep convolutional neural network for classifying images. *International Journal of Scientific and Research Publications*, 9(10): 143–150, 2019.

Yue Tan, Guodong Long, Jie Ma, Lu Liu, Tianyi Zhou, and Jing Jiang. Federated learning from pre-trained models: A contrastive learning approach. In *Proceedings of the 36th International Conference on Neural Information Processing Systems (NeurIPS)*, 2022.

Petroc Taylor. Uk households: Ownership of mobile telephones 1996-2022, Sep 21 2023. URL https://www.statista.com/statistics/289167/mobile-phone-penetration-in-the-uk/.

Connie W Tsao, Aaron W Aday, Zaid I Almarzooq, Cheryl AM Anderson, Pankaj Arora, Christy L Avery, Carissa M Baker-Smith, Andrea Z Beaton, Amelia K Boehme, Alfred E Buxton, et al. Heart disease and stroke statistics—2023 update: A report from the american heart association. *Circulation*, 147(8):93–621, 2023.

Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific Data*, 5 (1):1–9, 2018.

Grigorios Tsoumakas and Ioannis Katakis. Multi-label classification: An overview. *International Journal of Data Warehousing and Mining*, 3(3):1–13, 2007.

Hugo C Turner, Nguyen Van Hao, Sophie Yacoub, David A Clifton, Guy E Thwaites, Arjen M Dondorp, C Louise Thwaites, Nguyen Van Vinh Chau, et al. Achieving affordable critical care in low-income and middle-income countries. *BMJ Global Health*, 4(3):1–10, 2019.

Dennis Ulmer. A survey on evidential deep learning for single-pass uncertainty estimation. *arXiv preprint arXiv:2110.03051*, 2021.

Muhammad Habib ur Rehman, Ahmed Mukhtar Dirir, Khaled Salah, Ernesto Damiani, and Davor Svetinovic. Trustfed: A framework for fair and trustworthy cross-device federated learning in iiot. *IEEE Transactions on Industrial Informatics*, 17(12):8485–8494, 2021.

Akhil Vaid, Suraj K Jaladanki, Jie Xu, Shelly Teng, Arvind Kumar, Samuel Lee, Sulaiman Somani, Ishan Paranjpe, Jessica K De Freitas, Tingyi Wanyan, et al. Federated learning of electronic health records to improve mortality prediction in hospitalized patients with COVID-19: Machine learning approach. *JMIR Medical Informatics*, 9(1):1–10, 2021.

Joost Van Amersfoort, Lewis Smith, Yee Whye Teh, and Yarin Gal. Uncertainty estimation using a single deep deterministic neural network. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, pages 9690–9700. PMLR, 2020.

Jesper E Van Engelen and Holger H Hoos. A survey on semi-supervised learning. *Machine Learning*, 109(2):373–440, 2020.

Jason Van Hulse, Taghi M Khoshgoftaar, and Amri Napolitano. Experimental perspectives on learning from imbalanced data. In *Proceedings of the 24th International Conference on Machine Learning (ICML)*, pages 935–942, 2007.

Pieter Van Molle, Tim Verbelen, Bert Vankeirsbilck, Jonas De Vylder, Bart Diricx, Tom Kimpe, Pieter Simoens, and Bart Dhoedt. Leveraging the bhattacharyya coefficient for uncertainty quantification in deep neural networks. *Neural Computing and Applications*, 33(1):10259–10275, 2021.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the 30th International Conference on Neural Information Processing Systems (NeurIPS)*, pages 6405–6416, 2017.

Praneeth Vepakomma, Abhishek Singh, Otkrist Gupta, and Ramesh Raskar. Nopeek: Information leakage reduction to share activations in distributed deep learning. In *Proceedings of*

the International Conference on Data Mining Workshops (ICDMW), pages 933–942. IEEE, 2020.

EMILY A. VOGELS. About one-in-five americans use a smart watch or fitness tracker, Jan 9 2020. URL https://www.pewresearch.org/short-reads/2020/01/09/about-one-in-five-americans-use-a-smart-watch-or-fitness-tracker/.

Aidmar Wainakh, Fabrizio Ventola, Till Müßig, Jens Keim, Carlos Garcia Cordero, Ephraim Zimmer, Tim Grube, Kristian Kersting, and Max Mühlhäuser. User-level label leakage from gradients in federated learning. In *Proceedings on Privacy Enhancing Technologies*, pages 227–244, 2022.

Fei Wang and Anita Preininger. Ai in health: State of the art, challenges, and future directions. *Yearbook of Medical Informatics*, 28(1):16–26, 2019.

Lixu Wang, Shichao Xu, Xiao Wang, and Qi Zhu. Addressing class imbalance in federated learning. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pages 10165–10173, 2021.

Shuo Wang, Leandro L Minku, Nitesh Chawla, and Xin Yao. Learning from data streams and class imbalance. *Connection Science*, 31(2):103–104, 2019.

Wenqi Wei, Ling Liu, Margaret Loper, Ka-Ho Chow, Mehmet Emre Gursoy, Stacey Truex, and Yanzhao Wu. A framework for evaluating gradient leakage attacks in federated learning. *arXiv preprint arXiv:2004.10397*, 2020.

Karl Weiss, Taghi M Khoshgoftaar, and DingDing Wang. A survey of transfer learning. *Journal of Big data*, 3(1):1–40, 2016.

Max Welling and Yee W Teh. Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th International Conference on Machine Learning (ICML)*, pages 681–688. Citeseer, 2011.

Chuhan Wu, Fangzhao Wu, Lingjuan Lyu, Yongfeng Huang, and Xing Xie. Communication-efficient federated learning via knowledge distillation. *Nature communications*, 13(1):2032, 2022.

Tong Xia, Jing Han, Lorena Qendro, Ting Dang, and Cecilia Mascolo. Uncertainty-aware COVID-19 detection from imbalanced sound data. In *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 216–220, 2021a.

Tong Xia, Junjie Lin, Yong Li, Jie Feng, Pan Hui, Funing Sun, Diansheng Guo, and Depeng

Jin. 3dgcn: 3-dimensional dynamic graph convolutional network for citywide crowd flow prediction. *ACM Transactions on Knowledge Discovery from Data*, 15(6):1–21, 2021b.

Tong Xia, Yunhan Qi, Jie Feng, Fengli Xu, Funing Sun, Diansheng Guo, and Yong Li. Attnmove: History enhanced trajectory recovery via attentional network. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 35, pages 4494–4502, 2021c.

Tong Xia, Dimitris Spathis, J Ch, Andreas Grammenos, Jing Han, Apinan Hasthanasombat, Erika Bondareva, Ting Dang, Andres Floto, Pietro Cicuta, et al. COVID-19 sounds: A large-scale audio dataset for digital respiratory screening. In *Proceedings of the 35th Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Neurips)*, pages 1–13, 2021d.

Tong Xia, Jing Han, and Cecilia Mascolo. Benchmarking uncertainty quantification on biosignal classification tasks under dataset shift. *Multimodal AI in Healthcare: A Paradigm Shift in Health Intelligence*, 1(1):347–359, 2022a.

Tong Xia, Jing Han, and Cecilia Mascolo. Exploring machine learning for audio-based respiratory condition screening: A concise review of databases, methods, and open issues. *Experimental Biology and Medicine*, 247(22):2053–2061, 2022b.

Tong Xia, Jing Han, Lorena Qendro, Ting Dang, and Cecilia Mascolo. Hybrid-edl: Improving evidential deep learning for uncertainty quantification on imbalanced data. In *Proceedings of the Workshop on Trustworthy and Socially Responsible Machine Learning (WTSRML)*, 2022c.

Tong Xia, Jing Han, Abhirup Ghosh, and Cecilia Mascolo. Cross-device federated learning for mobile health diagnostics: A first study on COVID-19 detection. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023a.

Tong Xia, Yong Li, Yunhan Qi, Jie Feng, Fengli Xu, Funing Sun, Diansheng Guo, and Depeng Jin. History-enhanced and uncertainty-aware trajectory recovery via attentive neural network. *ACM Transactions on Knowledge Discovery from Data*, 2023b.

Tong Xia, Ting Dang, Jing Han, Lorena Qendro, and Cecilia Mascolo. Uncertainty-aware health diagnostics via class-balanced evidential deep learning. *IEEE Journal of Biomedical and Health Informatics*, 5(1):245–256, 2024a.

Tong Xia, Abhirup Ghosh, and Cecilia Mascolo. Flea: Improving federated learning on scarce and label-skewed data via privacy-preserving feature augmentation. *arXiv preprint arXiv:2312.02327*, 2024b.

Zhaohan Xiong, Martin K Stiles, and Jichao Zhao. Robust ECG signal classification for detection of atrial fibrillation using a novel neural network. In *Proceedings of the Computing in Cardiology (CinC)*, pages 1–4. IEEE, 2017.

Jenny Yang, Andrew AS Soltan, David W Eyre, and David A Clifton. Algorithmic fairness and bias mitigation for clinical machine learning with deep reinforcement learning. *Nature Machine Intelligence*, 1(1):1–11, 2023a.

Jenny Yang, Andrew AS Soltan, David W Eyre, Yang Yang, and David A Clifton. An adversarial training framework for mitigating algorithmic biases in clinical machine learning. *NPJ Digital Medicine*, 6(55):1–10, 2023b.

Qian Yang, Jianyi Zhang, Weituo Hao, Gregory P Spell, and Lawrence Carin. Flop: Federated learning on medical datasets using partial networks. In *Proceedings of the ACM Conference on Knowledge Discovery & Data Mining (KDD)*, pages 3845–3853, 2021.

Shan Yang, Heng Xiang, Qingda Kong, and Chunli Wang. Multi-label classification of electrocardiogram with modified residual networks. In *Proceedings of the Computing in Cardiology (CinC)*, pages 1–4. IEEE, 2020.

Wenti Yang, Naiyu Wang, Zhitao Guan, Longfei Wu, Xiaojiang Du, and Mohsen Guizani. A practical cross-device federated learning framework over 5g networks. *IEEE Wireless Communications*, 29(6):128–134, 2022a.

Xiangli Yang, Zixing Song, Irwin King, and Zenglin Xu. A survey on deep semi-supervised learning. *IEEE Transactions on Knowledge and Data Engineering*, 35(9):8934–8954, 2022b.

Yousef Yeganeh, Azade Farshad, Nassir Navab, and Shadi Albarqouni. Inverse distance aggregation for federated learning with non-iid data. In *Proceedings of the Workshop on Domain Adaptation and Representation Transfer, and Distributed and Collaborative Learning (MICCAI)*, pages 150–159. Springer, 2020.

Sofia Yfantidou, Dimitris Spathis, Marios Constantinides, Tong Xia, and Niels Van Berkel. Faircomp: Workshop on fairness and robustness in machine learning for ubiquitous computing. In *Adjunct Proceedings of the 2023 ACM International Joint Conference on Pervasive and Ubiquitous Computing & the 2023 ACM International Symposium on Wearable Computing*, pages 777–783, 2023.

Tehrim Yoon, Sumin Shin, Sung Ju Hwang, and Eunho Yang. Fedmix: Approximation of mixup under mean augmented federated learning. In *Proceedings of the International Conference on Learning Representations (ICLR)*, pages 1–10, 2020.

William J Youden. Index for rating diagnostic tests. *Cancer*, 3(1):32–35, 1950.

Dong Yu and Jinyu Li. Recent progresses in deep learning based acoustic models. *IEEE/CAA Journal of Automatica Sinica*, 4(3):396–409, 2017.

Felix Yu, Ankit Singh Rawat, Aditya Menon, and Sanjiv Kumar. Federated learning with only positive labels. In *Proceedings of the 30th International Conference on Machine Learning (ICML)*, pages 10946–10956. PMLR, 2020.

Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. Mixup: Beyond empirical risk minimization. In *Proceedings of the International Conference on Learning Representations (ICLR)*, pages 1–10, 2018.

Jie Zhang, Zhiqi Li, Bo Li, Jianghe Xu, Shuang Wu, Shouhong Ding, and Chao Wu. Federated learning with label distribution skew via logits calibration. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, pages 26311–26329. PMLR, 2022.

Kai Zhang, Jun Yu, Zhiling Yan, Yixin Liu, Eashan Adhikarla, Sunyang Fu, Xun Chen, Chen Chen, Yuyin Zhou, Xiang Li, et al. Biomedgpt: A unified and generalist biomedical generative pre-trained transformer for vision, language, and multimodal tasks. *arXiv preprint arXiv:2305.17100*, 2023a.

Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Fei Wu, et al. Instruction tuning for large language models: A survey. *arXiv preprint arXiv:2308.10792*, 2023b.

Xinwei Zhang, Mingyi Hong, Sairaj Dhople, Wotao Yin, and Yang Liu. Fedpd: A federated learning framework with adaptivity to non-iid data. *IEEE Transactions on Signal Processing*, 69(1):6055–6070, 2021.

Yifan Zhang, Bingyi Kang, Bryan Hooi, Shuicheng Yan, and Jiashi Feng. Deep long-tailed learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9): 10660–10671, 2023c.

Yuwei Zhang, Tong Xia, Abhirup Ghosh, and Cecilia Mascolo. Uncertainty quantification in federated learning for heterogeneous health data. In *Proceedings of the International Workshop on Federated Learning for Distributed Data Mining (FLDDM)*, pages 1–10, 2023d.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 2023.

Yue Zhao, Meng Li, Liangzhen Lai, Naveen Suda, Damon Civin, and Vikas Chandra. Federated learning with non-iid data. *arXiv preprint arXiv:1806.00582*, 2018.

Yuchao Zheng, Chen Li, Xiaomin Zhou, Haoyuan Chen, Hao Xu, Yixin Li, Haiqing Zhang,

Xiaoyan Li, Hongzan Sun, Xinyu Huang, et al. Application of transfer learning and ensemble learning in image-level classification for breast histopathology. *Intelligent Medicine*, 3(2): 115–128, 2023.

Zhi-Hua Zhou. *Ensemble methods: Foundations and algorithms*. CRC press, 2012.

Hangyu Zhu, Jinjin Xu, Shiqing Liu, and Yaochu Jin. Federated learning on non-iid data: A survey. *Neurocomputing*, 465(1):371–390, 2021.

Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He. A comprehensive survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 109(1):43–76, 2020.

Zefang Zong, Tong Xia, Meng Zheng, and Yong Li. Reinforcement learning for solving multiple vehicle routing problem with time window. *ACM Transactions on Intelligent Systems and Technology*, 1(1):1–10, 2024.

Quan Zou, Sifa Xie, Ziyu Lin, Meihong Wu, and Ying Ju. Finding the best classification threshold in imbalanced classification. *Big Data Research*, 5(1):2–8, 2016.