UNIVERSITY OF CALIFORNIA,
IRVINE

Towards Secure and Adaptive Healthcare AI: Intelligent Monitoring and Data Privacy

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Computer Science

by

Ziyu Wang

Dissertation Committee:
Professor Amir Rahmani, Chair
Distinguished Professor Nikil Dutt
Professor Ian G. Harris

2026

DEDICATION

To my family, friends, and mentors.

TABLE OF CONTENTS

Page

LIST OF FIGURES **vii**

LIST OF TABLES **ix**

ACKNOWLEDGMENTS **xi**

VITA **xv**

ABSTRACT OF THE DISSERTATION **xvii**

1 Introduction **1**
1.1 The Dual Imperative of Healthcare AI 1
1.2 From Adaptive Interaction to Privacy Auditing 3
1.2.1 Adaptive interaction and privacy-preserving learning 3
1.2.2 From interpretable ECG features to learned identity representations . 4
1.2.3 From re-identification capability to deployment-level privacy auditing 5
1.3 Research Questions 6
1.4 Contributions 7
1.5 Dissertation Organization 8

2 Context-Aware Active Stress Monitoring **9**
2.1 Introduction 9
2.2 Related Work 14
2.3 Offline Study 17
2.3.1 Proposed System Architecture 19
2.3.2 Sensor layer 19
2.3.3 Edge layer 20
2.3.4 Cloud layer 20
2.3.5 Preprocessing 21
2.3.6 Context-aware Active Reinforcement Learning Algorithm 23
2.4 Online Study 28
2.4.1 Proposed System Architecture 29
2.4.2 Preprocessing 31
2.5 Evaluation and Results 34
2.5.1 Offline Study: Evaluation and Results 35
2.5.2 Online Study: Evaluation and Results 39
2.5.3 Personalization Performance 41
2.5.4 Comparative Evaluation with Baseline Models 43
2.5.5 Ablation Study on System Components 44

2.5.6 Energy and Transmission Overhead Analysis . . . 46
2.6 Discussion . . . 47
2.6.1 System Limitations and Real-World Generalizability . . . 47
2.6.2 Privacy Considerations . . . 48
2.6.3 Reward Function Design and Policy Interpretability . . . 48
2.6.4 User-Centered Evaluation and Engagement . . . 49
2.6.5 Stress Modeling: Binary vs. Continuous . . . 49
2.6.6 Statistical Reliability and Class Imbalance . . . 49
2.6.7 Model Comparisons and Baselines . . . 50
2.6.8 Computational Overhead and Energy Usage . . . 50
2.6.9 Future Directions . . . 51
2.7 Conclusions . . . 51

3 Privacy-Preserving Adaptive Mental Health Monitoring 53
3.1 Introduction . . . 53
3.2 Methodology . . . 55
3.2.1 Differential Private Federated Transfer Learning Framework for Mental Health Monitoring . . . 56
3.2.2 Case Study on Stress Detection . . . 58
3.3 Evaluation . . . 60
3.4 Discussion . . . 62
3.5 Conclusion . . . 63

4 Interpretable ECG Re-identification Risk Analysis 64
4.1 Introduction . . . 64
4.2 Analysis of Re-identification Risks . . . 67
4.2.1 Method . . . 67
4.2.2 Experiment . . . 70
4.2.3 Results . . . 71
4.3 Conclusion . . . 74

5 Transformer-Based ECG Identity Leakage Analysis 75
5.1 Introduction . . . 75
5.2 Method . . . 77
5.2.1 Patch-based Modeling of Long-duration ECG . . . 77
5.2.2 Attention-based Identity Pattern Analysis . . . 79
5.3 Experimental Results . . . 79
5.3.1 Dataset and Experimental Setup . . . 79
5.3.2 Baselines . . . 80
5.3.3 Accuracy Assessment . . . 80
5.3.4 Attention-based Explainability . . . 81
5.4 Discussion . . . 83
5.5 Conclusion . . . 84

6 Linkage Attacks on Public ECG Data 85

6.1 Introduction . . . 85
6.2 Methodology . . . 88
6.2.1 Problem Definition . . . 89
6.2.2 Framework for Linkage Attack Evaluation . . . 90
6.2.3 ViT-Based Representation Learning . . . 90
6.2.4 Two-Stage Classification for Known and Unknown Participants . . . 91
6.2.5 Design Considerations and Practical Constraints . . . 93
6.3 Experiments and Results . . . 93
6.3.1 Datasets . . . 94
6.3.2 Experimental Setup . . . 94
6.3.3 Results and Analysis . . . 98
6.4 Discussion and Conclusion . . . 103

7 Participation Privacy in ECG Encoders **104**
7.1 Introduction . . . 104
7.2 Background and Related Work . . . 108
7.3 Threat Model and Problem Definition . . . 111
7.4 ECG Foundation Encoder Pretraining . . . 113
7.5 Membership Inference for Participation Privacy . . . 116
7.6 Experimental Setup . . . 121
7.6.1 Datasets and preprocessing . . . 121
7.6.2 Foundation encoder pretraining . . . 122
7.6.3 Membership inference evaluation . . . 122
7.7 Results . . . 124
7.7.1 Cohort Structure Drives Scalar Leakage . . . 125
7.7.2 The Value of Learning: When Learned Attackers Help (and When They Hurt) . . . 126
7.7.3 Embedding-Attack Results: Latent Geometry Can Be the Dominant Leakage Channel . . . 127
7.7.4 Calibration and Tail Risk at Low False-Positive Rates . . . 128
7.7.5 Summary of Empirical Principles . . . 129
7.8 Discussion and Limitations . . . 129
7.9 Conclusion . . . 131

8 Summary and Future Directions **137**
8.1 Summary of the Dissertation . . . 137
8.2 Cross-Cutting Findings . . . 139
8.3 Future Research Directions . . . 140
8.3.1 Human-Centered Adaptive Monitoring . . . 140
8.3.2 End-to-End Privacy Accounting . . . 140
8.3.3 Utility-Aware Representation Protection . . . 141
8.3.4 Longitudinal and Cross-Device Evaluation . . . 141
8.3.5 Causal and Clinically Grounded Explanations . . . 142
8.3.6 Adaptive Auditing and Deployment Controls . . . 142
8.3.7 Governance for Physiological Data and Models . . . 142

8.4 Concluding Perspective 143

Bibliography **144**

LIST OF FIGURES

Page

2.1 System Architecture - Offline Study 18
2.2 System Architecture - Online Study 30
2.3 Distribution of Stress Labels 34
2.4 Number of queries needed to reach a certain performance level during personalization. 37
2.5 Personalization Recall in Previous Work 38
2.6 Personalization ROC Curves 43
3.1 Overview of the proposed framework. 55
3.2 Overview of the proposed approach. 61
4.1 Overview of the threat model illustrating ECG data aggregation from various sources to e-health platforms, creating an attack surface for potential client re-identification. 65
4.2 ECG signal with detected PQRST peaks and key features, highlighting its biometric properties. 68
4.3 SHAP analysis for re-identification tasks: (a) gender, (b) age group, and (c) participant ID. 73
5.1 ECG data support digital-health applications but can also leak biometric identity. TransECG probes how transformer models encode identity-related ECG regions. 76
5.2 Overview of TransECG. Filtered ECG sequences are divided into temporal patches, encoded by a ViT-style transformer, and classified with a task-specific head. Final-layer class-token attention is mapped back to ECG morphology for interpretation. 77
5.3 ROC curves for (a) gender classification, (b) age classification, and (c) participant ID re-identification. 82
5.4 Representative attention maps for (a) gender classification, (b) age classification, and (c) participant re-identification, showing task-specific ECG patterns emphasized by the model. 82
5.5 ECG regions consistently emphasized across gender, age, and participant ID re-identification tasks. 83
6.1 Illustration of a linkage attack on public anonymous ECG data. A telehealth platform integrates data from multiple sources, including wearable devices and medical institutions. An attacker with partial knowledge cross-references public and private datasets to re-identify individuals, compromising privacy and trust in medical data sharing. 86
6.2 Overview of the proposed linkage attack method. 88
7.1 Motivation and threat surface in connected health. A foundation encoder trained on biosignal cohorts may be deployed through exposed interfaces, enabling membership inference and participation privacy leakage. 107

7.2 Participation privacy threat model in a connected-health ecosystem. A private patient cohort ($\mathcal{S}_{\text{train}}$) is aggregated to train a foundation ECG encoder (f_θ), which is subsequently deployed through a utility interface. An adversary probes the deployed encoder via observable outputs (e.g., scalar scores or aggregated statistics) and auxiliary public data ($\mathcal{S}_{\text{aux}}$) to infer whether a target subject contributed to pretraining. 110
7.3 Impact of learned attackers over score-only attacks in single-dataset pretraining. Each cell reports $\Delta\text{AUC} = \text{AUC}_{\text{learned}} - \text{AUC}_{\text{score}}$. Positive values indicate that training a learned attacker improves separability; negative values indicate that an objective-aligned score-only rule is stronger. 135
7.4 Learned-attack AUC across datasets and encoder families for single-dataset pretraining. The dashed line indicates chance-level AUC of 0.5. Leakage varies substantially across objectives and datasets, and the most severe cases occur in small-cohort settings. 136

LIST OF TABLES

Page

2.1 **Comparison of our study vs existing works** 17
2.2 **AWARE Features** 32
2.3 **PPG Features** 33
2.4 Classification performance results (mean ± SD) across participants using 5-fold cross-validation. 41
2.5 Personalization Results 43
2.6 Performance comparison for binary stress detection (Offline setting) using classical and deep learning models. Only full-feature models are shown. . . . 44
2.7 Performance comparison for binary stress detection (Online setting) using classical and deep learning models. All-feature models outperform HRV-only models. 44
2.8 Incremental evaluation of measured online stress-monitoring configurations using a Random Forest classifier. The final row reports the complete system and does not isolate the effects of RL triggering and personalization. 45
2.9 Estimated Daily Energy Consumption for Baseline Method 46
3.1 Stress Detection Performance of Different Models 61
4.1 Summary of ECG Datasets Used in Experiments 70
4.2 Data Splitting Methods for Re-identification Tasks 70
4.3 Performance Metrics for Re-identification Models 72
5.1 Performance comparison between TransECG and baselines. 81
6.1 Summary of ECG Datasets Used in Experiments 94
6.2 Baseline Comparison on Sample-Level Metrics. 99
6.3 Performance Under Different Data Splits (Partial Knowledge). 99
6.4 Performance Under Adversarial Scenarios. 100
6.5 Misclassification Rates of Linkage Attack. 101
6.6 Confidence-Based Misclassification in Linkage Attack. 102
7.1 **Diverse real-world ECG corpora used in this study.** We report (i) *intrinsic dataset properties* (cohort, acquisition setting, original sampling rate, available leads, and typical record duration) from the official PhysioNet releases, and (ii) *the effective scale used in our workspace* after our deterministic preprocessing. **Preprocessing (shared across all datasets):** we resample all signals to 250 Hz, extract 10-second windows with 5-second stride, select a single ECG channel (prefer lead II when available; otherwise use an available ECG lead), remove non-finite values, and apply per-record z-normalization before caching windows to disk. These datasets jointly cover heterogeneous acquisition devices, cohort sizes, and recording durations, reflecting practical connected-health conditions rather than a single curated benchmark. 132

7.2 **Learned membership inference.** Results represent `[AUC|TPR@0.01|Adv]` evaluated under a **mixed-dataset non-member setting**. Bold indicates the highest leakage for each dataset. 133
7.3 **Score-only membership inference.** Comparison of leakage using only scalar outputs. Small non-zero artifacts in TPR are corrected to 0.000, and negative advantage scores are clipped to 0.000. Bold indicates the highest leakage for each dataset. 133
7.4 Membership inference with embedding access 134

ACKNOWLEDGMENTS

There are many people I would like to thank for being part of my Ph.D. journey. I have never been particularly good at expressing my feelings, nor do I usually think of myself as an emotional person. Completing this dissertation, however, gives me a rare opportunity to say what I have not always managed to say. Along the way, I received more support and kindness than I could ever fully repay. I hope that, in the years ahead, I can carry that kindness forward and, in my own way, help light the way for others, just as so many people have done for me. I have been incredibly fortunate, and this journey has been more meaningful than I could have imagined.

First and foremost, I would like to express my deepest gratitude to my advisor and committee chair, Professor Amir M. Rahmani. The first two months of my Ph.D. were not easy. In November 2022, he welcomed me into the UCI Health SciTech Group and placed his trust in my potential. I began my doctoral studies determined to find my place in AI for healthcare and to make a meaningful contribution to the field. Amir gave me the opportunity and freedom to pursue that goal. I wanted to prove, both to him and to myself, that his decision had been the right one. Today, as I complete this dissertation, I can finally say: yes, I did it. I am deeply grateful for his guidance, encouragement, trust, and support throughout this journey.

I am also sincerely grateful to Distinguished Professor Nikil Dutt for his encouragement and support at several important moments during my Ph.D., as well as for serving on both my advancement and final defense committees. I thank Professor Ian G. Harris for also serving on these committees and for his thoughtful feedback and valuable suggestions. I am equally grateful to Professors Ramesh Jain and Salma Elmalaki for serving on my advancement examination committee and for their support and insightful feedback.

During my four years with the UCI Health SciTech Group, I grew tremendously as both a researcher and a person. I was fortunate to meet so many people who became not only collaborators and colleagues, but also close friends. I am especially grateful to Elahe Khatibi. We worked side by side on many papers, shared the rewards and frustrations of research, and supported each other through the difficulties and joys of both work and life. More importantly, we became close friends who could always speak honestly about our careers, our lives, and everything in between.

I would also like to thank Seyed Amir Hossein Aqajari, who mentored me on my first project in the group. He patiently taught me how to write pandas code line by line. That was my first real lesson in machine learning after moving from computer systems security into the field, and I have never forgotten it. I am equally grateful to Zhongqi Yang, Yong Huang, Iman Azimi, Nitish Nagesh, Hamidreza Alikhani Koshkak, Pengfei Zhang, Yutong Song, Mahyar Abbasian, Milad Asgari, Salar Jafarlou, Sina Labbaf, Mahkameh Rasouli, Saba A. Farahani, Arshia Ilaty, and Chenhan Lyu. If I have inadvertently left anyone out, please forgive me; your presence and support have been no less important to me.

Beyond the Health SciTech Group, I have had the privilege of working with many outstanding collaborators. I am grateful to Professor Sanaz Rahimi Moosavi, Professor Farshad Firouzi, Dr. Anil Kanduri, Professor Marco Levorato, Professor Krishnendu Chakrabarty, Professor Pasi Liljeberg, Dr. Ankita Sharma, and Dr. Kianoosh Kazemi for their collaboration, expertise, and contributions to my research.

The text of Chapter 2 of this dissertation is a reprint of the material described in the following publication: S. A. H. Aqajari, Z. Wang, A. Tazarv, S. Labbaf, S. Jafarlou, B. Nguyen, N. Dutt, M. Levorato, and A. M. Rahmani. "Enhancing Performance and User Engagement in Everyday Stress Monitoring: A Context-Aware Active Reinforcement Learning Approach." Accepted for publication in *ACM Transactions on Computing for Healthcare (HEALTH)*, 2026. Seyed Amir Hossein Aqajari and Ziyu Wang contributed equally to this work. The coauthors listed in this publication are Seyed Amir Hossein Aqajari, Ali Tazarv, Sina Labbaf, Salar Jafarlou, Brenda Nguyen, Nikil Dutt, Marco Levorato, and Amir M. Rahmani. Amir M. Rahmani directed and supervised the research that forms the basis of this chapter. Copyright for this material is held by the authors, and publication rights are licensed to ACM.

The text of Chapter 3 of this dissertation is a reprint of the material described in the following publication: Z. Wang, Z. Yang, I. Azimi, and A. M. Rahmani. "Differential Private Federated Transfer Learning for Mental Health Monitoring in Everyday Settings: A Case Study on Stress Detection." In *2024 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 5402–5406, 2024. Ziyu Wang and Zhongqi Yang contributed equally to this work. The coauthors listed in this publication are Zhongqi Yang, Iman Azimi, and Amir M. Rahmani. Amir M. Rahmani directed and supervised the research that forms the basis of this chapter. This material is reprinted with permission from IEEE, which holds the copyright.

The text of Chapter 4 of this dissertation is a reprint of the material described in the following publication: Z. Wang, A. Kanduri, S. A. H. Aqajari, S. Jafarlou, S. R. Mousavi, P. Liljeberg, S. Malik, and A. M. Rahmani. "ECG Unveiled: Analysis of Client Re-identification Risks in Real-World ECG Datasets." In *2024 IEEE 20th International Conference on Body Sensor Networks (BSN)*, pages 1–4, 2024. The coauthors listed in this publication are Anil Kanduri, Seyed Amir Hossein Aqajari, Salar Jafarlou, Sanaz R. Mousavi, Pasi Liljeberg, Shaista Malik, and Amir M. Rahmani. Amir M. Rahmani directed and supervised the research that forms the basis of this chapter. This material is reprinted with permission from IEEE, which holds the copyright.

The text of Chapter 5 of this dissertation is based on material accepted for publication by IEEE in the following work: Z. Wang, E. Khatibi, K. Kazemi, I. Azimi, S. Mousavi, S. Malik, and A. M. Rahmani. "TransECG: Interpreting Biometric Identity Leakage in ECG Signals via Vision Transformers." In *22nd IEEE-EMBS International Conference on Body Sensor Networks (BSN 2026)*. IEEE, 2026. Ziyu Wang and Elahe Khatibi contributed equally to this work. The coauthors of this work are Elahe Khatibi, Kianoosh Kazemi, Iman Azimi, Sanaz Mousavi, Shaista Malik, and Amir M. Rahmani. Amir M. Rahmani directed and

supervised the research that forms the basis of this chapter. This material is used with permission from IEEE, which holds the copyright.

The text of Chapter 6 of this dissertation is a reprint of the material described in the following publication: Z. Wang, E. Khatibi, F. Firouzi, S. R. Mousavi, K. Chakrabarty, and A. M. Rahmani. "Linkage Attacks Expose Identity Risks in Public ECG Data Sharing." In *2025 47th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 1–7, 2025. Ziyu Wang and Elahe Khatibi contributed equally to this work. The coauthors listed in this publication are Elahe Khatibi, Farshad Firouzi, Sanaz Rahimi Mousavi, Krishnendu Chakrabarty, and Amir M. Rahmani. Amir M. Rahmani directed and supervised the research that forms the basis of this chapter. This material is reprinted with permission from IEEE, which holds the copyright.

The text of Chapter 7 of this dissertation is a reprint of the material described in the following publication: Z. Wang, E. Khatibi, A. Sharma, K. Chakrabarty, S. R. Moosavi, F. Firouzi, and A. Rahmani. "Membership Inference Attacks Expose Participation Privacy in ECG Foundation Encoders." *Smart Health*, 41:100680, 2026. Ziyu Wang and Elahe Khatibi contributed equally to this work. The coauthors listed in this publication are Elahe Khatibi, Ankita Sharma, Krishnendu Chakrabarty, Sanaz Rahimi Moosavi, Farshad Firouzi, and Amir Rahmani. Amir Rahmani directed and supervised the research that forms the basis of this chapter. This material is reprinted with permission from Elsevier Inc., which holds the copyright.

Financial support for portions of this research was provided by the U.S. National Science Foundation, Division of Computer and Network Systems, under Secure and Trustworthy Cyberspace Grant CNS-2344869, and by the UC Noyce Initiative Grant.

I am deeply grateful to my friends Weiyu Chen, Hanning Chen, Wenjun Huang, Wenjie Qu, Hao Li, Di Huang, Elahe Khatibi, Zhuoyuan Li, Pengfei Zhang, Yutong Song, Zhongqi Yang, Yuqiao Li, Monica Zhou, Jiacheng Liang, Hanzhi Li, Yu Duan, Yirui He, Zixia Xia, Ismat Jarin, Boning Yang. Thank you for your friendship and support, and for sharing with me the many moments of a Ph.D. journey that was not always easy, but was deeply rewarding. I hope each of you finds success, fulfillment, and happiness in the years ahead. To the many friends whose names are not listed here, please know that you are no less important to me and that you have always had a place in my heart.

I would also like to thank badminton and all my badminton friends. Badminton has been a passion of mine for fifteen years and stayed with me throughout my entire Ph.D. journey. Whenever research became stressful, the court was where I could clear my mind, enjoy the moment, and feel like myself again. Thank you for every game, every laugh, and every friendship along the way.

I would also like to express my sincere gratitude to Ms. Gan for her support, understanding, and encouragement during an important stage of my Ph.D. journey. Her kindness and companionship made those years more meaningful, and I will always be grateful.

Finally, my deepest gratitude goes to my parents, my grandmothers, my aunt and uncle, and the rest of my family for their unconditional love and support. Their love and care have been a constant source of strength throughout my life. Everything I have accomplished rests, in no small part, on all that they have given me. More than anything, I wish them a lifetime of health and happiness.

VITA

Ziyu Wang

EDUCATION

Doctor of Philosophy in Computer Science **2026**
University of California, Irvine *Irvine, California*

Master of Science in Computer Science **2025**
University of California, Irvine *Irvine, California*

Bachelor of Engineering in Information Security **2022**
Huazhong University of Science and Technology *Wuhan, China*

FIELD OF STUDY

Computer Science

RESEARCH EXPERIENCE

Graduate Student Researcher **2022–2026**
University of California, Irvine *Irvine, California*

INDUSTRY EXPERIENCE

Machine Learning Engineer Intern **2026**
Apple *Santa Clara, California*

Applied Scientist Intern **2025**
Amazon *Seattle, Washington*

PUBLICATIONS INCLUDED IN THIS DISSERTATION

S. A. H. Aqajari, Z. Wang, A. Tazarv, S. Labbaf, S. Jafarlou, B. Nguyen, N. Dutt, M. Levorato, and A. M. Rahmani. "Enhancing Performance and User Engagement in Everyday Stress Monitoring: A Context-Aware Active Reinforcement Learning Approach." Accepted for publication in *ACM Transactions on Computing for Healthcare (HEALTH)*, 2026.

Z. Wang, Z. Yang, I. Azimi, and A. M. Rahmani. "Differential Private Federated Transfer Learning for Mental Health Monitoring in Everyday Settings: A Case Study on Stress Detection." In *2024 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 5402–5406, 2024.

Z. Wang, A. Kanduri, S. A. H. Aqajari, S. Jafarlou, S. R. Mousavi, P. Liljeberg, S. Malik, and

A. M. Rahmani. "ECG Unveiled: Analysis of Client Re-identification Risks in Real-World ECG Datasets." In *2024 IEEE 20th International Conference on Body Sensor Networks (BSN)*, pages 1–4, 2024.

Z. Wang, E. Khatibi, K. Kazemi, I. Azimi, S. Mousavi, S. Malik, and A. M. Rahmani. "TransECG: Interpreting Biometric Identity Leakage in ECG Signals via Vision Transformers." In *22nd IEEE-EMBS International Conference on Body Sensor Networks (BSN 2026)*. IEEE, 2026.

Z. Wang, E. Khatibi, F. Firouzi, S. R. Mousavi, K. Chakrabarty, and A. M. Rahmani. "Linkage Attacks Expose Identity Risks in Public ECG Data Sharing." In *2025 47th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 1–7, 2025.

Z. Wang, E. Khatibi, A. Sharma, K. Chakrabarty, S. R. Moosavi, F. Firouzi, and A. Rahmani. "Membership Inference Attacks Expose Participation Privacy in ECG Foundation Encoders." *Smart Health*, 41:100680, 2026.

ABSTRACT OF THE DISSERTATION

Towards Secure and Adaptive Healthcare AI: Intelligent Monitoring and Data Privacy

By

Ziyu Wang

Doctor of Philosophy in Computer Science

University of California, Irvine, 2026

Professor Amir Rahmani, Chair

Artificial intelligence is transforming healthcare from episodic assessment into continuous and personalized monitoring. In practice, a monitoring system rarely ends at the sensor: physiological measurements, contextual observations, model updates, or learned representations may pass from a wearable device to a phone, an edge gateway, a cloud service, or an institutional platform for storage, collaborative learning, clinical interpretation, and model reuse. These flows make intelligent monitoring scalable and adaptive, but they also expand the privacy surface beyond conventional identifiers. This dissertation develops a continuous research program around this operational tension: it begins with user-aware data acquisition, adds privacy-preserving adaptive learning, and then progressively examines what sensitive information can remain in the data and artifacts that move through the healthcare AI lifecycle.

The first part of the dissertation focuses on adaptive health monitoring. A context-aware active reinforcement-learning framework learns when to request ecological momentary assessments by considering prediction uncertainty together with time, response history, and smartphone context. The framework progresses from offline label selection to real-time deployment and personalized stress detection, showing how a monitoring system can collect

informative labels while reducing unnecessary user burden. Once these data are collected, limited on-device samples and the need to improve models across users motivate collaborative learning infrastructure. A differential private federated transfer-learning framework therefore combines public-data pretraining, decentralized user adaptation, and protected aggregation. This approach supports learning from sparse and distributed health data while reducing the need to centralize raw observations and limiting disclosure through shared model updates.

Federated learning and protected aggregation reduce exposure at a specific stage, but they do not eliminate the broader need to share, process, and reuse healthcare information. Raw records may later be de-identified for research, representations may be transferred to downstream applications, and pretrained encoders may be deployed through cloud or institutional interfaces. The second part therefore follows the residual privacy risk as information changes form along this pipeline. An interpretable analysis of electrocardiogram (ECG) data demonstrates that physiological morphology itself can reveal demographic attributes and participant identity across heterogeneous datasets. TransECG extends this analysis to transformer representations and uses model-intrinsic attention to localize identity-related evidence within physiologically meaningful ECG regions. The dissertation then moves from measuring biometric information to evaluating its operational consequences: a linkage attack shows how de-identified ECG records can be connected across data sources under partial adversarial knowledge, while a membership-inference audit shows that reusable ECG foundation encoders can reveal whether an individual or cohort participated in pretraining through exposed scores or embeddings. Together, these studies demonstrate that removing explicit identifiers or decentralizing training does not make the remaining lifecycle artifacts inherently anonymous: privacy leakage can arise from the signal itself, from learned representations, from cross-dataset sharing, and from deployed model interfaces.

The central conclusion is that secure healthcare intelligence must be adaptive at two levels: it must adapt interaction and learning to the user, and it must adapt privacy safeguards to

the changing attack surface across the data and model lifecycle. No single mechanism, such as de-identification or decentralized training, is sufficient on its own. Trustworthy healthcare AI instead requires burden-aware sensing, privacy-preserving learning, interpretable risk analysis, realistic adversarial evaluation, and continuous auditing of the interfaces through which data and models are shared.

Chapter 1

Introduction

1.1 The Dual Imperative of Healthcare AI

Intelligent health systems are moving from episodic clinical measurements toward continuous, data-driven monitoring. Wearable devices, smartphones, and connected medical sensors can observe health outside the clinic, while machine-learning models convert these observations into personalized assessments and opportunities for earlier intervention. This transition is especially valuable in settings such as mental-health monitoring, where physiological and contextual signals vary substantially across people and labeled data are often sparse. The same opportunities and constraints extend to high-dimensional clinical imaging and biomechanical measurements. In ophthalmic applications, semi-supervised learning can reduce expert annotation for retinal ganglion cell identification in three-dimensional adaptive optics optical coherence tomography (AO-OCT), while unified learning frameworks can jointly detect cone photoreceptors and measure inner- and outer-segment lengths from AO-OCT B-scans [155, 156]. Machine-learning systems have also been used to detect keratoconus from Ocular Response Analyzer waveforms, estimate ocular stiffness through adaptive temporal

mixtures of experts, and automate the montaging of multi-gaze ultrasound scans for measuring ocular globe shape [154, 152, 153]. These examples illustrate a broader healthcare AI challenge: useful models must learn from complex, specialized, and often sparsely labeled measurements while remaining reliable across people and acquisition settings. A practical monitoring service, however, rarely consists of an isolated sensor and a model running permanently on one device. Measurements or derived artifacts commonly move through a pipeline that can include a wearable, a smartphone, an edge gateway, cloud infrastructure, a clinical platform, and a research repository. Storage, cross-user learning, remote analysis, longitudinal review, and reuse of pretrained models all depend on some form of communication across this pipeline.

The movement of information is not incidental to intelligent monitoring; it is part of how monitoring becomes scalable, adaptive, and clinically useful. It is also how a local measurement acquires a broader privacy surface. Raw signals may be replaced by gradients, embeddings, predictions, or model parameters, and explicit identifiers may be removed before records are shared, but each artifact can preserve information about the person from whom it originated. Federated learning can reduce raw-data centralization, and differential privacy can bound particular disclosures during training, yet neither mechanism automatically anonymizes a physiological signal, a released dataset, a learned representation, or a deployed model interface. Privacy must therefore be evaluated not only where data are collected, but also whenever information is transformed, transmitted, aggregated, released, or queried.

This dissertation begins from the premise that adaptivity and privacy are connected design objectives across this full lifecycle. An adaptive model benefits from learning individual structure, whereas a private system must control how that structure is exposed as data and models move between operational contexts. The research develops this argument in three phases that follow the path of healthcare information. It first makes everyday monitoring

responsive to user context and supports protected learning from sparse, distributed observations. It then investigates the residual identity information contained in physiological signals and increasingly expressive representations. Finally, it tests whether that information can be exploited through shared datasets and reusable foundation encoders under realistic adversarial conditions. The resulting progression connects collection, decentralized training, data sharing, representation learning, and deployment as stages of one healthcare AI system rather than as separate technical problems.

1.2 From Adaptive Interaction to Privacy Auditing

The technical progression of the dissertation follows information through a real-world health-monitoring system. Each phase begins with an operational need—acquiring timely labels, learning across users, sharing data for research, or reusing trained models—and then examines the privacy consequences of the artifact that enables it. This organization makes the transition from monitoring to privacy analysis explicit: the data flows required for useful monitoring are also the pathways through which sensitive information may be exposed.

1.2.1 Adaptive interaction and privacy-preserving learning

The first phase asks how a monitoring system can obtain useful supervision and adapt to individuals without imposing unnecessary burden or unnecessarily centralizing sensitive observations. Continuous health monitoring depends on labels that connect sensor measurements to an individual's lived experience, yet ecological momentary assessments become intrusive when they are frequent or poorly timed. Chapter 2 addresses this problem with a context-aware active reinforcement-learning framework that uses prediction uncertainty, time, response history, and smartphone context to decide when to request stress labels [7].

The work progresses from an offline analysis of query efficiency to an online deployment that triggers assessments in real time and supports personalized stress detection. The resulting labels remain limited, imbalanced, and privacy-sensitive. At the same time, learning a model that generalizes beyond one device or one person requires knowledge to be aggregated across users, often with coordination by cloud or institutional infrastructure. Chapter 3 develops differential private federated transfer learning for this setting [141]. A model is pretrained with a larger nonsensitive source dataset, adapted through federated fine-tuning, and protected by noise applied to aggregated updates. Taken together, these chapters establish two complementary forms of adaptation: the system adapts its interaction to the monitored person and its model to distributed user data, while reducing raw-data movement and protecting the updates needed for collaborative learning.

1.2.2 From interpretable ECG features to learned identity representations

Federated and differentially private training address how a model learns from distributed records, but their guarantees are scoped to that learning procedure. A functioning healthcare ecosystem may still retain signals for longitudinal analysis, release de-identified datasets for reproducible research, transfer embeddings to downstream tasks, or provide access to pretrained encoders. The second phase therefore asks what privacy-sensitive information survives as monitoring data are transformed and reused beyond their original collection context. Electrocardiogram (ECG) is a particularly important case because its morphology reflects anatomy, cardiac conduction, demographic factors, and stable individual physiology. Chapter 4 uses transparent machine-learning models and SHAP analysis across heterogeneous real-world datasets to identify the PQRST features that support gender, age-group, and participant re-identification [134]. This analysis establishes that ECG privacy risk originates in the signal itself, is interpretable, and is not confined to opaque deep networks. Hand-

crafted features, however, reveal only part of the privacy surface because modern representation learners can integrate information across much longer temporal contexts. Chapter 5 consequently introduces TransECG, a Vision Transformer-based framework that improves re-identification performance while mapping model attention back to physiologically meaningful ECG regions [136]. Its attention analysis shows that identity-related evidence is distributed across depolarization, conduction, and repolarization components. The progression from interpretable features to transformer representations exposes a central dual-use property of healthcare AI: the representations that improve analytical utility and enable model reuse can also preserve and amplify latent identity structure.

1.2.3 From re-identification capability to deployment-level privacy auditing

Demonstrating that a signal contains identity information does not by itself establish whether that information is exploitable under realistic adversarial constraints. The third phase therefore moves from re-identification capability to operational privacy threats. Closed-set identity classification assumes that every target belongs to a known class, which overstates an attacker's knowledge and poorly represents open data ecosystems. Chapter 6 relaxes this assumption through a linkage attack in which an adversary has only partial auxiliary knowledge and must distinguish known from unknown individuals [135]. A two-stage Vision Transformer attack with a dynamic confidence threshold shows that public ECG records can still be linked to identities across heterogeneous sources. The final step extends the system boundary from released data to reusable models. In the setting studied in Chapter 7, neither raw waveforms nor identity labels need to be disclosed; an adversary instead queries scalar scores or latent representations and asks whether a person or cohort participated in pretraining. The chapter evaluates score-only, learned, and embedding-based membership inference attacks across contrastive and masked-reconstruction objectives [138]. Its results reveal het-

erogeneous, objective-dependent leakage, with particularly substantial participation risk for small or institution-specific cohorts. These two chapters show that privacy auditing must account for both the ways data are shared and the behavior of the model interfaces through which learned representations are deployed.

1.3 Research Questions

This dissertation addresses six research questions:

1. How can context, response history, prediction uncertainty, and reinforcement learning be combined to request informative stress labels while limiting user burden?

2. How can transfer learning, federated learning, and differential privacy be integrated to support personalized health monitoring under limited and imbalanced user data?

3. Which interpretable ECG features expose demographic and participant identity across heterogeneous real-world datasets?

4. How do transformer representations encode identity-related information over long-duration ECG signals, and which physiological regions drive their decisions?

5. How well can an adversary link anonymized ECG records under partial-knowledge and open-set conditions?

6. Do reusable ECG foundation encoders reveal whether an individual or cohort participated in pretraining through score or embedding interfaces?

Together, these questions connect human-aware data acquisition and privacy-preserving system design with adversarial evaluation. They follow the same information as its operational role expands: from interactions with the monitored individual, to learning across distributed

local records, to the reuse of shared signals and representations, to cross-dataset identity linkage, and finally to participation leakage from deployed models.

1.4 Contributions

The principal contributions of this dissertation are:

- A context-aware active reinforcement-learning framework that schedules ecological momentary assessments using user context and response history, progresses from offline label selection to online triggering, and evaluates personalization in everyday stress monitoring.
- A differential private federated transfer learning framework that combines public-data pretraining, decentralized user adaptation, and protected aggregation for data-scarce mental-health monitoring.
- An interpretable, multi-dataset analysis of ECG re-identification that quantifies gender, age-group, and participant identity risks and traces them to clinically meaningful PQRST features.
- TransECG, an attention-based transformer framework for studying how long-duration ECG representations encode biometric identity beyond handcrafted features.
- A realistic ECG linkage-attack formulation that incorporates partial adversarial knowledge, unknown participants, heterogeneous datasets, and confidence-aware identity matching.
- A subject-centric membership-inference audit of self-supervised ECG foundation encoders across model objectives, datasets, attacker knowledge, and deployment inter-

faces.

- A unified account of healthcare intelligence as a lifecycle property, showing that user-aware adaptation and privacy protection must begin during data acquisition and continue through learning, sharing, and deployment.

1.5 Dissertation Organization

Chapter 2 presents context-aware, online label acquisition and personalized stress monitoring. Chapter 3 builds on this setting with privacy-preserving adaptive learning from sparse, distributed data. Chapter 4 establishes the biometric privacy risk of shared ECG data using interpretable models. Chapter 5 examines how transformer models capture and localize identity information. Chapter 6 converts re-identification capability into a realistic cross-dataset linkage attack. Chapter 7 extends the privacy analysis to participation inference against reusable ECG foundation encoders. Finally, Chapter 8 synthesizes the findings and outlines future directions for adaptive, privacy-aware healthcare AI.

Chapter 2

Enhancing Performance and User Engagement in Everyday Stress Monitoring: A Context-Aware Active Reinforcement Learning Approach

This chapter is based on material published in ACM Transactions on Computing for Healthcare (HEALTH).

2.1 Introduction

As per data from the American Institute of Stress [6], approximately 55% of individuals in the United States encounter stress throughout their day. The American population stands as one of the most stressed globally, with their current stress levels surpassing the global average by 20 percentage points. The impact of stress extends to the physical body, cognitive processes,

emotions, and behavior [80, 137]. Unaddressed stress can contribute to several health issues, including high blood pressure, heart ailments, obesity, and diabetes [80]. Consequently, daily life monitoring of stress has garnered significant significance within our society, and the advancement of techniques for diagnosing human stress holds paramount importance [3, 4].

The presence of stress within the human body can be diagnosed by analyzing psychophysiological signals such as Photoplethysmography (PPG) and Electrocardiogram (ECG) [22, 70, 134]. PPG is a simple optical sensing technique for detecting blood volume alterations within peripheral circulation [5, 138]. With the rapid advancements in technology and the development of the Internet of Things (IoT) [147, 140, 67, 2], the acquisition of PPG signals has been greatly facilitated, primarily through the utilization of wearable devices like smart rings or smartwatches [20, 135]. Consequently, monitoring stress levels in everyday life becomes an attainable endeavor by analyzing the PPG signals garnered from the aforementioned wearable devices. Furthermore, the evolution of mobile apps designed for context logging has provided a means to consistently observe and record a user's contextual data, including elements like their location, activities, weather conditions, and other relevant variables, all in real-time [104]. Prior studies have already demonstrated the significance of this contextual information in understanding and identifying stressful experiences encountered by individuals [119, 17, 59]. In everyday situations, biosignals vary widely among individuals due to physiological and lifestyle differences, as well as the diverse activities one might partake in. Additionally, the perception of stress levels differs from person to person, leading to biases in the data collected. Therefore, there is a significant need to tailor predictive models to each individual across their various activities. These adjustments, which are essential at the time of deployment, present both conceptual and technical challenges.

The evolution of stress monitoring in daily settings has seen a significant transformation. Originally, stress monitoring was largely confined to controlled environments such as laboratories, which allowed researchers to closely observe and study physiological responses

under stress [49, 118]. Such controlled studies laid the groundwork for understanding stress responses in a structured setting. Over time, advancements in technology enabled the transition to more naturalistic and dynamic environments. This shift paved the way for methods like Ecological Momentary Assessments (EMAs) [16]. EMAs revolutionized stress monitoring by allowing real-time data collection about participants' stress levels throughout their day-to-day activities. Using self-reported stress (EMA) as the labeling source, users are asked to respond to real-time queries that link the data gathered by sensors to stress labels in everyday situations. One of the main challenges is optimally collecting these labels (stress levels) from individuals in their daily lives [77]. Frequently triggering EMAs or sending them at inappropriate times, such as when a user is busy with work or sleeping, could burden the user. This could lead to a significantly lower number of reported labels. Moreover, selecting the optimal moments to send the EMAs, especially during instances when an individual is experiencing stress, poses a considerable challenge and holds the utmost importance.

To tackle these challenges, in this work, we introduced a contextual variant of active learning, based on Deep Q-Learning, which incorporates the contextual information pertaining to an individual into the decision-making process. In the initial phase of our research, a context-aware active reinforcement learning algorithm was utilized in an offline setting [126]. This approach was implemented to thoroughly evaluate the effectiveness of our proposed method. We demonstrated that the utilization of such an algorithm in a stress detection task can lead to a reduction of up to 88% in the required EMAs when compared to a random selection approach, and up to 32% when compared to traditional active learning methods. Furthermore, we observed that employing such an algorithm can increase the performance of stress detection tasks by up to 21% compared to a random selection method, and up to 8% when compared to traditional active learning approaches. However, an offline context-aware active reinforcement learning algorithm abstains from employing active learning to initiate EMAs. However, this approach may still entail user burdens and result in triggering EMAs at inappropriate times.

In this article, an extension of our previous work [126], we improve the proposed algorithm for application in an online setting by using active learning to initiate EMAs. In the online setting, the algorithm initiates EMAs at different points during the study, guided by contextual information about the user. The active learning algorithm analyzes this information in real time to determine appropriate moments for requesting stress labels. This adaptive approach is designed to reduce query burden while improving label utility. We evaluate the offline and online variants in two longitudinal study cohorts collected during different periods and compare their performance using identical classification algorithms and evaluation metrics. The results show higher stress-detection performance for the online cohort and motivate the use of context-aware, real-time EMA triggering. Lastly, we employ a personalization technique to investigate the effects of personalized customization in enhancing the model's performance.

Novelty Beyond Prior Offline Work. This manuscript extends our prior offline context-aware active reinforcement learning study [126] in several important directions. First, instead of performing label selection only after data collection, we deploy the RL policy *online* to decide whether to trigger an EMA in real time, enabling adaptive label acquisition under realistic user constraints (e.g., availability and burden). Second, we integrate real-time contextual signals and user response history into the decision loop, which was not supported in the offline setting. Third, we conduct a new longitudinal, IRB-approved online study (March 2022–May 2023) and evaluate the end-to-end system with real-time sensing, transmission, and EMA triggering. Finally, we systematically investigate personalization effects in the online pipeline, demonstrating how subject-specific fine-tuning improves stress prediction robustness in everyday settings.

In summary, the key contributions of this paper are as follows:

- Propose a new form of active learning, utilizing Deep Q-learning, aimed at enhancing

interaction with the monitored individual during data collection.

- Develop a sensor-edge-cloud layered system architecture for the acquisition and labeling of the data aimed at real-time stress detection. We further demonstrate the effectiveness of our proposed system through the utilization of actual real-time data and comparing it with an identical offline variant of the algorithm.

- Incorporate the contextual features into the stress detection models for the purpose of systematically monitoring the influence of contextual factors within the context of stress detection.

- Examine how the performance of our algorithm is enhanced with the inclusion of subject-specific data during the training phase in order to explore the influence of personalization on stress detection.

- Conduct a two-stage IRB-approved study on 54 individuals across undergraduate and graduate student populations over two periods: June 2020 to June 2021 (offline method) and March 2022 to May 2023 (online method), generating a total of 132,598 filtered samples.

This paper is structured as follows: Section 2 offers a comprehensive review of stress assessment methods and associated research, highlighting the importance of personalizing models and the crucial role of user behavior and contextual data in real-time labeling. It also emphasizes the need to enhance user engagement in everyday settings. Section 3 details the platform developed for data collection and analysis. Section 4 outlines the dataset we gathered and our data processing methods. Section 5 introduces our proposed context-aware active learning approach, which incorporates user behavior and context into its query mechanism, along with temporal data correlations, to enhance query scheduling. Section 6 reports on the outcomes of various querying techniques in relation to personalization. Finally, Section 7 concludes the paper and includes a discussion.

2.2 Related Work

Stress-related research often delineates its origins from both exogenous factors—such as lifestyle, interpersonal relationships, and financial stability—and endogenous factors like individual psychological constitution and thought processes. These factors act as progenitors for negative affective states, including anxiety and fear, and instigate corresponding physiological responses. The physiological aspect of stress, denoted as a stress response, is a series of bodily reactions to environmental stimuli or stressors. Within the scientific discourse, the construct of stress is categorized into psychological, behavioral, and physiological dimensions. Historically, self-report measures such as the Perceived Stress Scale (PSS), formulated by Cohen et al. [32], and the stress inventory by Holmes and Rahe [64], have been the standard for gauging stress levels retrospectively.

Nonetheless, the accuracy of survey-based assessments of stress is compromised by measurement biases, including response bias, which reflects the influence of the query's framing on the participant's responses. Additionally, while some behavioral expressions of stress—like facial expressions—are spontaneous, they may also be subject to volitional control, thus potentially skewing data accuracy. Consequently, recordings of such behaviors must be critically examined for systematic errors that may misrepresent the actual stress magnitude.

Given these constraints, and paralleling the evolution of high-precision sensor technology, there is an augmented demand for veritable detectors of physiological stress markers. Biosignal attributes of stress episodes are typically involuntary, and such data can be acquired through methodologies like electrocardiography (ECG), photoplethysmography (PPG), electromyography (EMG), skin conductance (SC) or electrodermal activity (EDA), respiratory rate (RSP), skin temperature (ST), pupil dilation (PD), and cerebral activity as captured by electroencephalography (EEG) [54].

The current methodologies for monitoring stress in daily life utilize EMAs to inquire about

participants' stress levels throughout the day [77]. The task of effectively gathering accurate stress level indicators from individuals in the context of their everyday activities presents a significant challenge [111]. The over-frequent activation of EMAs or their issuance at times that clash with a user's schedule, such as during work hours or rest periods, can be an imposition. This may culminate in a reduced quantity of reported stress labels. Additionally, pinpointing the precise moments for sending EMAs, particularly in moments when stress levels are elevated, is of paramount significance and presents a notable challenge.

Existing methods in the literature for the deployment of EMAs in daily life stress monitoring studies can be categorized into three distinct categories: 1. Random, 2. Time-based, and 3. Statistical-based.

Within the random triggering methods, EMAs are dispatched at random intervals throughout the course of the study. Random sampling is often the preferred approach in situations where the research topic's indicators cannot be reliably ascertained [35]. However, when the research topic possesses specific objectives and focus, alternative methods tend to yield more favorable outcomes.

Time-based triggering methods involve sending EMAs at fixed pre-defined intervals throughout the day. The majority of existing research endeavors in daily life stress monitoring in the literature employ this algorithm for their label querying system, as documented in previous studies [149, 150, 86, 133, 12]. While this algorithm boasts simplicity of implementation and uniform coverage of the study period, it imposes a considerable burden on participants due to untimely EMA deliveries. Consequently, this may result in an increased prevalence of missing data and a reduction in the utility of collected labels.

In the context of statistical-based triggering methods, EMAs are dispatched based on the distribution of samples [125]. Under this triggering algorithm, a label is requested for a specific sample based on the number of unlabeled samples within its vicinity. Although this

policy can effectively mitigate the incidence of undesired EMA deliveries, it may still impose a burden on users and lead to missing data, as it fails to consider contextual information about users, which is pivotal in determining the optimal times for EMA deployment.

In this study, we propose a context-aware active reinforcement learning approach to effectively trigger EMAs throughout the day.

Initially we conducted an offline study where we employed a statistical-based triggering method to send EMAs throughout the day. In this phase, the probability of selecting each sample for labeling is proportionate to the quantity of prior unlabeled samples in its proximity. This approach increases the likelihood of requesting a user label for a sample situated in a region with a substantial number of unlabeled samples. Upon accumulating a sufficient number of labeled samples for each region, the data collection process is terminated. We implemented three distinct algorithms offline for optimal label selection in model development: 1. Random, 2. Traditional Active Reinforcement Learning, and our novel approach 3. Context-Aware Active Reinforcement Learning. Our findings revealed that using a context-aware active reinforcement learning algorithm in stress detection significantly decreases the necessity for EMAs enhances the effectiveness of stress detection over random or traditional active learning methodologies. As previously noted, statistical-based triggering algorithm may still impose a user burden and result in missing data due to its failure to incorporate user contextual information into the label-querying decision-making process.

In the next phase, we propose an online context-aware active reinforcement learning algorithm to utilize RL agent for decision making in real time to further improve the performance. This algorithm actively utilizes a context-aware active learning approach in real-time based on Deep Q-Learning to determine whether an EMA should be triggered for a given sample. By considering real-time contextual information related to each user in the decision-making process of whether to trigger an EMA for a specific sample, our approach is poised to significantly reduce the user burden associated with untimely EMA deliveries, consequently

leading to an increase in the acquisition of high-utility labels. Table 2.1 provides a summary of the comparison between our work (offline and online studies) and existing literature on this subject.

Table 2.1: **Comparison of our study vs existing works**

Study	**Triggering Method**	**EMA Frequency**	**Real-time Analysis**	**Data-based queries**	**Context-based queries**	**Online triggering method**
Yu et al. [149]	Time-based	1.5 hours apart, 10/day	No	No	No	No
Yu et al. [150]	Time-based	1/day	No	No	No	No
Mundnich et al. [86]	Time-based	1/day	No	No	No	No
Wang et al. [133]	Time-based	Every three months	No	No	No	No
Battalio et al. [12]	Random and Time-based	End of the day, varies	Yes	Yes	No	No
Our Offline Study	Statistical-based	A cap of seven EMAs per day	Yes	Yes	No	No
Our Online Study	**Active Reinforcement Learning**	**A cap of seven EMAs per day**	Yes	Yes	Yes	Yes

2.3 Offline Study

Data Flow and Outputs. In the offline study, the system processes PPG signals collected from a smartwatch and optionally includes contextual features from the smartphone (e.g., screen status, call activity, and time of day). The PPG signals are preprocessed using the HeartPy library to extract heart rate variability features, which, along with contextual features, serve as inputs to a machine learning classifier (e.g., Random Forest). The output

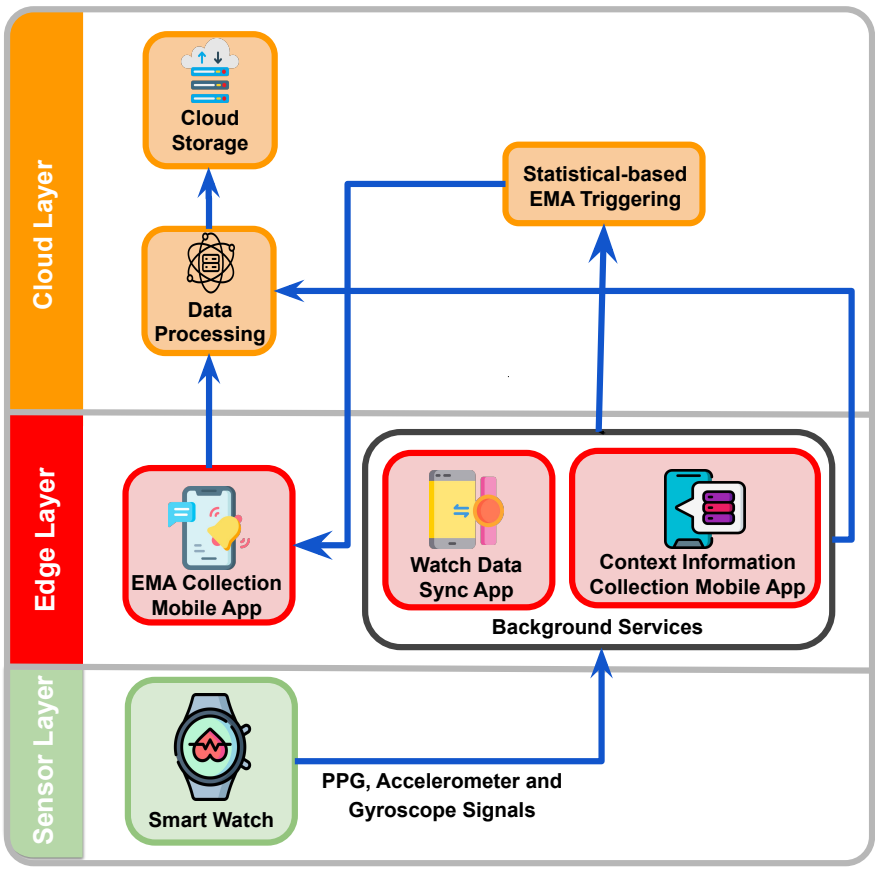


Figure 2.1: System Architecture - Offline Study

is a binary stress label (stressed or not stressed), inferred using supervised learning based on user-annotated EMA responses.

In the initial stage of this work, we target to evaluate the effectiveness of our proposed label triggering method [126]. The EMAs are dispatched to participants' phones on a statistical-based basis. Once data collection was completed, we applied our proposed method of context-aware active reinforcement learning for the labeling process. Our offline study involved an Institutional Review Board (IRB)-sanctioned study on human subjects, during which we gathered over 2,629 days of data in everyday environments from college students.

The collected dataset encompasses PPG and motion measurements, including acceleration, gyroscope, and gravity readings, and is partially annotated with stress levels, emotional states, and physical activities obtained through EMAs delivered at statistically determined

intervals.

Our data collection initiative, spanning from June 2020 to June 2021, involved 20 volunteers selected from undergraduate and graduate student populations. The demographic breakdown of the participants included 13 male and 7 female students, ranging in age from 19 to 29 years. During the study, we gathered 109,586 samples over a period of 2,629 days. Participants contributed to the study for periods ranging from 11 to 287 days, with an average participation duration of 131 days. On average, each subject contributed 5,479 instances to the dataset.

2.3.1 Proposed System Architecture

Creating a reliable system to gather real-time physiological and contextual data while using active learning from participants is challenging. Wearable devices like smartwatches can be affected by motion artifacts, requiring extensive processing for stress detection [110]. Timing label requests is crucial to ensure participant engagement and label reliability. Figure 2.1 depicts our offline proposed three-layer system including a sensor layer for data acquisition, an edge layer for data transmission and user interaction, and a cloud layer for data processing and decision-making respectively. This system architecture illustrates the architectural composition of our proposed three-layer system, called ZotCare [115].

2.3.2 Sensor layer

This study employs Samsung Galaxy Gear Sport Watches as the designated wearable devices. These smartwatches are outfitted with sensors capable of recording PPG (20Hz), accelerometer, and gyroscope (motion) signals [106]. To facilitate the acquisition of these unprocessed PPG and motion signals, we have developed a smartwatch application tailored for Samsung

Galaxy Gear Sport Watches operating on the Tizen operating system [130]. When the watch is connected to a local Wi-Fi network, it efficiently transmits the amassed data to the cloud layer. In situations where a local Wi-Fi network is unavailable, the data is transmitted via Bluetooth to a synchronized smartphone. The raw signal acquisition program comprises two distinct services and a user interface (UI). The initial service manages the periodic delivery of sensor data to the cloud, with data transmissions occurring every fifteen minutes and data intervals set at two-minute increments.

2.3.3 Edge layer

We employ the AWARE framework [9] to collect contextual data in everyday scenarios. AWARE is an open-source mobile tool designed for recording, sharing, and reusing context-related information on mobile devices. It utilizes the built-in sensors of smartphones to capture various aspects of daily life, including battery status, weather conditions, location, screen activity, and more. In situations where Wi-Fi connectivity is unavailable, we utilize an alternative smartphone application installed on the edge of our network. This application collects raw PPG signals and accelerometer data from the sensor layer through Bluetooth and subsequently transmits this data to cloud storage. To obtain stress level ratings from our study participants, we have developed an additional smartphone application. This application employs an EMA approach to request stress level assessments from the participants.

2.3.4 Cloud layer

This layer comprises two distinct modules:

- **Data Processing:** This module focuses on processing data retrieved from the edge layer, with its primary objectives being encryption and the storage of data in the

appropriate format on ZotCare servers.

- **Statistical-based EMA Triggering:** Our label triggering method is consists of two phases:

  - **Initial Stage:** To obtain an initial approximation of the sample distribution within the sample space, we start the procedure with observation. During the first N samples (100 samples in our configuration, equivalent to approximately 25 hours of wearing the watch), no EMAs are initiated. By the conclusion of this phase, an estimation of the sample distribution in the sample space is obtained.

  - **Query Stage:** Subsequent to the initial phase (from N+1 onward), EMAs are triggered (labels requested) for a subset of samples. The selection probability for labeling each sample is proportionate to the number of preceding samples (unlabeled) in its proximity. This approach ensures that samples in regions with a substantial number of unlabeled counterparts are more likely to be queried for labels. Once a sufficient number of labeled samples are acquired for a particular region, the label collection for that region ceases. Nonetheless, the minimum probability of triggering an EMA for a sample is set at P = 0.1. Consequently, even if a sample is situated in an area with few or no previous samples, the probability of a query remains nonzero. This design enables exploration of unseen regions as well as regions with higher densities while maintaining a balanced approach.

2.3.5 Preprocessing

Following the conclusion of study and data collection, the collected raw PPG signals are preprocessed in order to extract relevant features for model building.

Data Cleaning

The bio-signals from wearable devices, which are inherently noisy, are stored directly in the cloud. The goal of this step is to remove clearly erroneous data points. We use the motion data to remove noises and artifacts in the bio-signals. In our study, we primarily focus on refining PPG signals and heart-rate data. For PPG signals, we utilize a bandpass Butterworth filter with cutoff frequencies between 0.7 Hz and 3.5 Hz. To ensure consistency, the filter is of the third order, and we adopt a sampling rate of 20 Hz, which aligns with our data collection parameters. Additionally, we implement a moving average over a 1-second window to smoothen the PPG data, mitigating artifacts commonly induced by body gestures and movements in daily environments.

Data Normalization

Normalization is indispensable when aiming to minimize variances specific to individual participants and countering the repercussions of subpar bio-signal samples. A standout method in this context, particularly in statistical analyses and machine learning, is the min-max normalization. This technique scales feature values consistently within a predetermined range, commonly [0, 1]. By doing so, it ensures the inherent structure of the data remains intact, which becomes crucial for algorithms that might be affected by the magnitude of the features. For our PPG bio-signals, we've employed the min-max normalization as our primary estimator to ensure that each feature aligns appropriately within a range dictated by the training set.

Feature Extraction

In our investigation, a feature extraction module was employed to analyze PPG data in 2-minute intervals. This facilitated the identification of PPG peaks and the derivation of key metrics such as heart rate. Utilizing the HeartPy library [128] for comprehensive PPG signal processing, we extracted 12 features from both electrical activations and pressure waveforms in the dataset. The extracted PPG features are presented in Table 2.3.

2.3.6 Context-aware Active Reinforcement Learning Algorithm

Our study utilizes the Context-aware Active Reinforcement Learning algorithm to label the collected data in our offline study. In this section, we provide a detailed explanation of this method and compare it with the random selection method and traditional active reinforcement learning methods for label querying.

In traditional supervised learning, the entire labeled dataset was utilized for training [74]. However, within our personalized data collection approach, we accumulated labeled data from diverse sources over time. To leverage this, we iteratively queried our users for new annotations. Commencing with a subset of labeled data, we employed a query mechanism to identify the most informative unlabeled instances with the assistance of human experts.

Active learning played a pivotal role in this process, strategically selecting data samples for labeling to enhance model accuracy while minimizing data usage. Strategies encompassed uncertainty sampling (opting for ambiguous data) and diversity sampling (choosing unique and indicative data). Nevertheless, real-world queries incurred costs, necessitating a delicate balance between query expenses and model improvement.

In our scenario, we encountered distinctive challenges. User behavior influenced data quality and availability, contingent on factors such as activity, time, query frequency, and phone

interaction. A lack of response resulted in the denial of labeling and delayed responses, diminishing data quality alignment. Consequently, our active learning approach needed to consider not only data quality but also future label accessibility. We proposed the use of Deep Q-Learning, where an agent modeled user behavior to ensure sustained user engagement.

Deep Q-learning

Why DQN? We adopt DQN rather than simpler heuristics or contextual bandits because EMA triggering is inherently sequential and exhibits delayed effects. Issuing an EMA at an inconvenient moment may reduce the participant's future responsiveness, while spacing queries appropriately can improve long-term engagement and label yield. Such long-horizon trade-offs are naturally modeled as a Markov Decision Process (MDP), where the optimal policy depends not only on the current stress uncertainty but also on temporal context (e.g., time of day and time since last query) and user-specific response dynamics. Contextual bandits optimize myopic reward without modeling state transitions, and rule-based heuristics require hand-crafted schedules that do not generalize well across participants or daily routines. DQN provides a principled framework to learn an adaptive policy that balances label utility and user burden over time.

Deep Q-Learning (DQN) [62] is a model-free reinforcement-learning method that can learn online from off-policy experience. At its core, DQN seeks to estimate the action-value function, denoted as $Q(s, a)$, which predicts the expected return after taking an action a in state s. The Bellman equation, which is fundamental to Q-learning, is given by:

$$Q(s, a) = r + \gamma \max_{a'} Q(s', a') \tag{2.1}$$

Where r is the immediate reward, γ is the discount factor, and s' is the subsequent state after taking action a in state s. The primary distinction between traditional Q-learning

and DQN is the utilization of deep neural networks to approximate the Q-values. This is paramount for tasks with large state spaces. The loss $\mathcal{L}$ during training is defined as:

$$\mathcal{L}(\theta) = \mathbb{E}_{(s,a,r,s')\sim U(D)} \left[\left(r + \gamma \max_{a'} Q(s', a'; \theta^-) - Q(s, a; \theta) \right)^2 \right] \tag{2.2}$$

Where D is the replay buffer, $U(D)$ is a uniform random sample from D, θ are the network parameters, and θ^- are the target network parameters.

In the following part, we will elaborate the detailed definitions in our DQN.

State Representation

The state vector, denoted by s, encodes crucial information leading to decision-making. The intention is to refine and personalize the stress detector to optimize accuracy while minimizing user queries. The state comprises:

- **Uncertainty Factor**: Originating from the raw output of a pre-trained classifier. This factor measures the distance from the decision boundary, effectively quantifying the confidence of the prediction.
- **Time-aware Response Rate**: This accounts for the time of the day (in hourly intervals) and embodies the user's responsiveness across different hours.
- **Time since Last Query**: To enhance user experience and prevent excessive querying in short time frames.
- **Time of Day**: Represents the current hour and is used to model potential variations in stress levels throughout the day.

Reward Formulation

The reward function integrates components from the ‘n_state‘ vector for holistic decision-making. It is formulated as:

$$r_0 = \frac{1}{1 + e^{-\alpha_0 (n_{\text{state[0]}} - 0.5)}} \tag{2.3}$$

$$r_1 = \text{reward\_F}(n_{\text{state[1]}}) \tag{2.4}$$

$$r_2 = \frac{1}{1 + e^{-\alpha_2 (n_{\text{state[2]}} - 0.5)}} \tag{2.5}$$

Here, α_0 and α_2 are hyperparameters controlling the slope of the sigmoid functions, tuned empirically during training to balance responsiveness and stability of the reward signals.

The overall reward function, R, based on the action taken, is:

$$R(\text{action}) = \begin{cases} \text{reward\_p} & \text{if action is True} \\ 3 - \text{reward\_p} & \text{if action is False} \end{cases} \tag{2.6}$$

Q-Network Design

The Q-network constitutes the backbone of our framework. The structure and features are enumerated below:

- The core is a densely connected neural network geared towards estimating Q-values.
- Input: The network takes in 4 nodes, matching the count of state variables.
- Hidden Layers: The architecture consists of variable hidden layers, as specified by the

list h. In the provided example, four hidden layers are employed with 5, 9, 7, and 5 nodes, respectively. Each of these nodes uses the ReLU activation function and incorporates both $l1$ and $l2$ kernel regularizers, with the $l2$ regularization strength set at $1e^{-2}$.

- Output: The network furnishes 2 output nodes, indicative of the duo of feasible actions, with a linear activation function.

Additionally, in our experiments, the agent employed an ϵ-greedy policy accompanied by a linear annealing schedule for its exploration factor. This strategy ensures a gradual transition from exploration to exploitation during the learning process, thereby enhancing convergence and robustness in diverse environments. For our implementation, we leveraged the Keras-RL library [96]. To elucidate, the ϵ-greedy policy in reinforcement learning can be characterized as follows:

$$\pi(a|s) = \begin{cases} \epsilon + \frac{1-\epsilon}{|A|} & \text{if } a = \text{argmax}_{a' \in A} Q(s, a') \\ \frac{1-\epsilon}{|A|} & \text{otherwise} \end{cases} \tag{2.7}$$

Where:

- $\pi(a|s)$ is the probability of taking action a in state s.
- ϵ is the exploration probability.
- A is the set of possible actions.
- $Q(s, a')$ is the estimated value of taking action a' in state s.

For our experiment, we employed a sequential memory architecture with a capacity limited to 50,000 instances and a window length set at one. The DQN agent was initialized with

parameters set as follows: a discount factor γ at 0.95, a warm-up phase consisting of 100 steps, and a learning rate of $1e^{-2}$ for the target model update. Optimization was carried out using the Adam optimizer, and the performance was gauged using the Mean Absolute Error (MAE) metric.

Policy Strategy

The decision strategy, symbolized by $\pi_\theta(s, a)$, selects the action that corresponds to the optimum Q-value for a specified state s. To guarantee a complete traversal of the state space:

- On an estimated 5% of occasions, a query is initiated randomly. This procedure considers the potentially undiscovered areas within the state space.
- Following every $K = 100$ occasions, which ideally occurs once in a 24-hour span, the user's response frequency metrics are re-calibrated according to their interaction patterns. This adjustment acknowledges individual variability and revitalizes the query selection methodology tailored for each user.
- The stress detection module undergoes periodic retraining. This process assimilates the most recent subjectively labeled information, amalgamated with the prior objective data procured from diverse subjects.

2.4 Online Study

Data Flow and Output. In the online deployment, real-time contextual features and user response history are used as inputs to an RL agent that decides whether to trigger an EMA at each decision interval. If triggered, the EMA collects a binary stress label from the user.

Recent PPG and context features are processed by the cloud-based stress-detection pipeline, while newly collected labels support scheduled personalization updates. The system outputs (1) EMA trigger decisions and (2) real-time stress predictions. We note that "online" in this work refers to real-time system deployment and decision making for EMA triggering, rather than continuous online model adaptation or streaming parameter updates.

In our online study, we use the same context-aware active-learning DQN (context-aware AL-DQN) examined offline. We evaluate it in a real-time setting where user responses inform the agent's decisions. By choosing when to request labels, this approach is designed to reduce unnecessary queries and improve stress-detection performance relative to the offline approach.

We assessed data from a cohort of 34 individuals. The study spanned from March 2022 to May 2023. Participants ranged in age from 19 to 29 years. After filtering anomalous and noisy records, we retained 23,012 samples collected over 420 participant-days, with an average of 676 samples per participant. The study protocol was approved by the responsible Institutional Review Board.

2.4.1 Proposed System Architecture

The proposed system architecture is illustrated in Figure 2.2. Consistent with our offline study, we maintain a three-layer system, denoted as ZotCare. A comparison with the architecture presented in Figure 2.1 reveals that, although the sensor layer and the edge layer remain unchanged, substantial modifications are implemented in the cloud layer. These alterations are undertaken to render the system conducive to real-time label querying through the utilization of our proposed context-aware active reinforcement learning algorithm explained in 2.3.6.

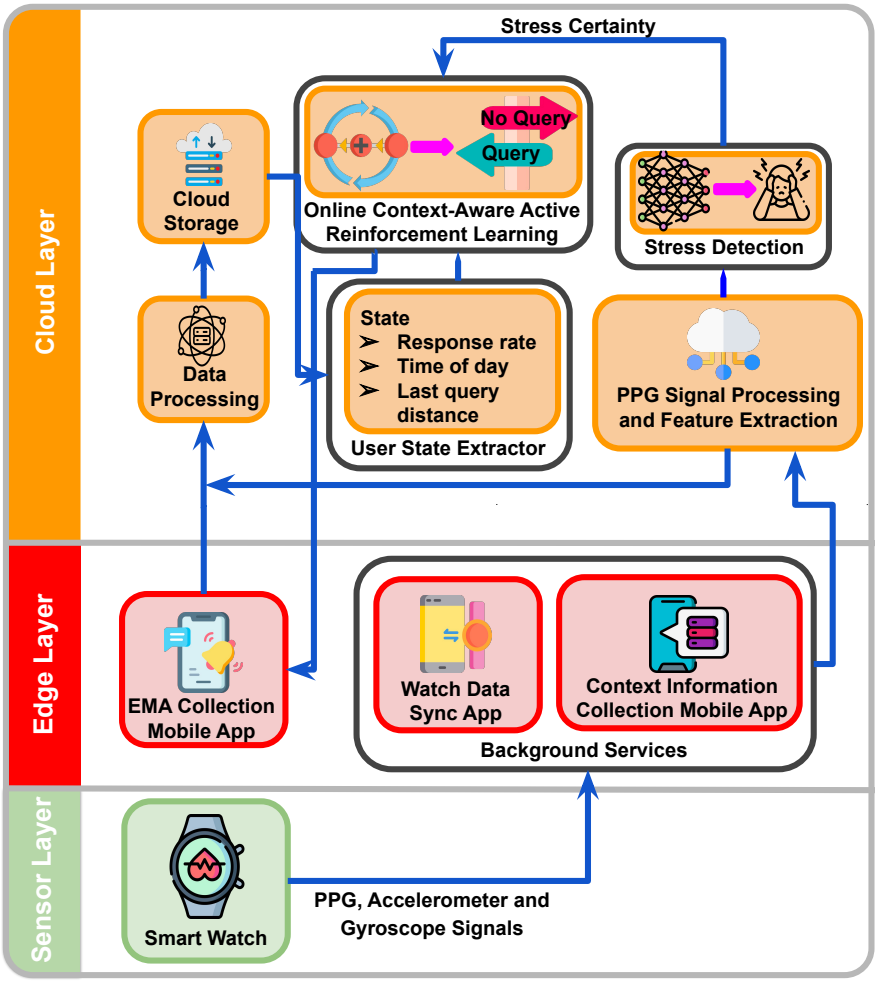


Figure 2.2: System Architecture - Online Study

The cloud layer comprises four modules that replace the statistical triggering method with the proposed context-aware triggering algorithm.

- PPG Signal Preprocessing and Feature Extraction: To effectively identify moments conducive to experiencing stress, we continuously monitor participants' stress levels in real time using PPG signals from their watches. However, to make these signals suitable for stress prediction, several preprocessing steps are required. This module is dedicated to preparing the PPG signals for stress detection. More details can be found in Section 2.4.2.

- Stress Detection: We utilize the data from our previous study [126] to construct our

stress detection module. The features extracted from the PPG signals are input into this module for stress detection. The level of certainty regarding stress is then forwarded to the context-aware active reinforcement learning module to aid in identifying stressful moments.

- Context Recognition: Within this module, we extract contextual information pertaining to each user, which is subsequently provided to our active learning module for decision-making purposes. This includes factors such as the time elapsed since the last query, the time of day, and the time-aware response rate. The time-aware response rate considers the user's responsiveness within the current hour based on their historical activity.

- Context-Aware Active Reinforcement Learning: The primary objective of this module is to determine whether it is appropriate to trigger an EMA at any given moment. Stress certainty, time elapsed since the last query, time of day, and time-aware response rate are all input into this module to inform the decision-making process. If an EMA needs to be triggered, a notification is dispatched to the user's mobile device on the edge layer, prompting them to rate their current stress level. In the following section, we will delve deeper into the training process of the active reinforcement learning agent.

2.4.2 Preprocessing

PPG signals, contextual AWARE data, and user-reported stress levels are collected from the cloud for stress model construction. However, raw cloud-stored PPG and AWARE data need preprocessing before building the model. This section explains our data preparation steps.

Data Cleaning and Normalization

This study employs the same modules for data cleaning and normalization as discussed in the offline study section (see Section 2.3.5).

Feature Extraction

- PPG Features: This module was discussed in Section 2.3.5; Table 2.3 lists the extracted PPG features.
- Contextual Features: The raw contextual information obtained from AWARE is not ready for building the stress detection models. We transform both categorical and numerical raw features into solely numerical features. We show the features extracted from raw AWARE data in Table 2.2.

Table 2.2: **AWARE Features**

Feature	**Definition**
Call	Call duration, type, and count
Notification	APP source and count
Screen & Touch	User screen interactions
Battery	Battery charge duration and level
Message	Message type and count
Time	Time of the day (24-hour format)
Location	Longitude, latitude, altitude

Data Labeling

The EMA protocol is configured to prompt participants up to seven times per day to report their perceived stress on a five-point scale: (1) not at all, (2) a little bit, (3) some, (4) a lot, and (5) extremely. Each EMA response is timestamped and stored in the cloud for subsequent alignment with sensor data.

Table 2.3: **PPG Features**

Feature	**Definition**
BPM	Heart beats per minute
IBI	Inter-Beat Interval, the average time interval between two successive heartbeats (NN intervals)
SDNN	Standard deviation of NN intervals
SDSD	Standard deviation of successive differences between adjacent NNs
RMSSD	Root mean square of successive differences between the adjacent NNs
PNN20	The proportion of successive NNs greater than 20ms
PNN50	The proportion of successive NNs greater than 50ms
HR_mad	Median absolute deviation of NN intervals
SD1 and SD2	Standard deviations of the corresponding Poincare plot
S	Area of ellipse described by SD1 and SD2
BR	The number of breaths per minute (breathing rate)

Physiological data, including raw PPG signals, is collected in 2-minute windows every 15 minutes. This periodic schedule balances energy efficiency with sufficient physiological resolution. After each collection, sensor-derived features and contextual indicators (e.g., screen activity, call frequency) are uploaded to the cloud on the same 15-minute cycle.

To assign stress labels to each 15-minute interval of sensor data, we implemented a nearest-neighbor alignment strategy: each interval is labeled based on the closest subsequent EMA response provided by the participant. This approach allows for temporally relevant labeling while preserving the naturalistic conditions of the study. Although this introduces a minor delay between data collection and labeling, it reflects the realistic constraints of real-world deployments and follows practices established in prior ubiquitous sensing research [86]. The resulting distribution of stress labels is illustrated in Figure 2.3.

Following common practice in everyday stress monitoring, we additionally report binary stress detection results by mapping EMA ratings of “a lot” and “extremely” to stressed and the remaining ratings to not stressed. This binarization mitigates sparsity at the highest stress levels and reduces subjectivity between adjacent categories, enabling more stable learning and evaluation. We acknowledge that the 15-minute alignment may introduce temporal

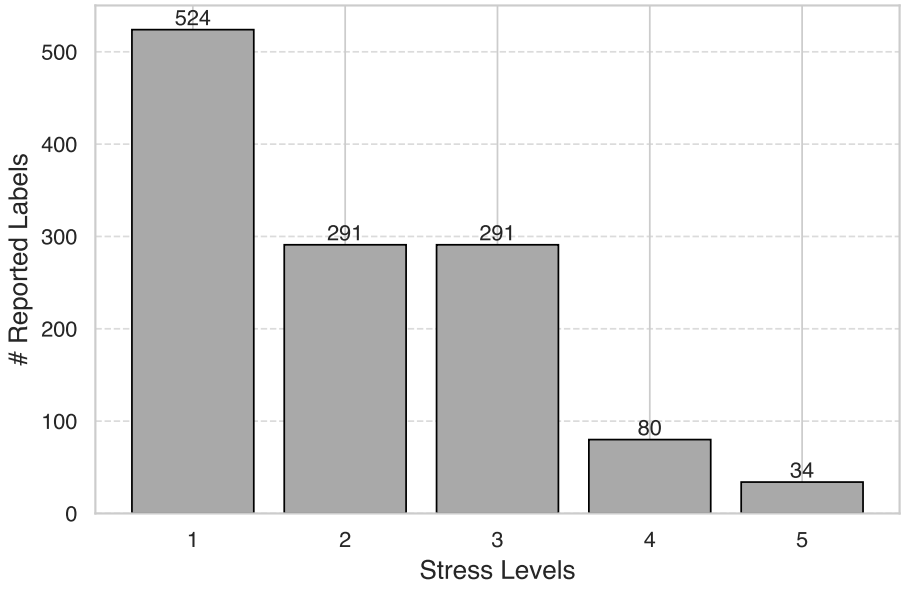


Figure 2.3: Distribution of Stress Labels

smoothing (e.g., if stress changes rapidly within the window); however, this design reflects realistic deployment constraints and prioritizes consistent, low-burden sensing and labeling.

2.5 Evaluation and Results

In this section, we present a comprehensive evaluation of our proposed context-aware active reinforcement learning approach, organized into two main parts: the *offline study* and the *online study*. This clear separation enables a more coherent understanding of each phase of the experiment, respective methods, and findings.

2.5.1 Offline Study: Evaluation and Results

Offline Model Pre-training and Setup

We initiated the pre-training of the stress detector, which was subsequently utilized to derive the state and reward in constructing the active learning framework. A random forest classifier with n = 500 estimators (number of trees) and max depth = 5 for each tree was employed. Participants were requested to assess their stress levels on a five-point scale: (1) not at all, (2) a little bit, (3) some, (4) a lot, and (5) extremely. We translated the stress labels into two categories: a lot and extremely as 1 (stressed), and the remaining three labels as 0 (not stressed). This binary formulation is consistent with prior ubiquitous stress monitoring work and helps reduce ambiguity across adjacent EMA categories while addressing imbalance in high-stress labels.

We conducted training for the model using labeled data from 14 different subjects, reserving one subject for personalization. The newly trained model on this subject exhibited a recall value of recall = 0.238 for the minority class (class stressed). The Q-learning agent underwent pre-training with an offline sequential objective dataset. Upon completing a sequence (referred to as one episode), we restarted from the beginning until reaching the total number of steps. The model underwent training for K=200,000 steps. The total reward achieved during the episodes reached a saturation point before this stage, indicating the model's convergence.

Training Modes of the Agent

The agent underwent training in two distinct modes. Initially, we employed a "traditional" approach, where the agent's attention was solely on the state and reward associated with the classifier's raw output. The agent received significant rewards for executing the 'submit

query' action for instances within the classifier's uncertainty region, while being rewarded for the 'do not submit query' action for other instances. In this mode, the agent operated without capturing contextual information and functioned akin to a conventional active learning selection policy.

Subsequently, a modification was introduced by incorporating contextual information into the reward function, as previously described. High rewards were assigned for the 'submit query' action for instances not only within the uncertainty region but also from a time interval when the user demonstrated increased responsiveness. Moreover, instances not in a short time distance from the preceding 'select' action were considered. For other instances, the agent received rewards for executing the 'do not submit query' action.

Query Efficiency Analysis

We conducted an analysis on the quantity of queries needed to achieve a particular level of personalization. The experiment was repeated N = 100 times to mitigate the influence of random selection. The average number of instances is presented in Figure 2.4. Notably, a substantial disparity exists between random selection and DQN agents, with the context-aware agent demonstrating the capability to attain high performance with a significantly lower number of queries. Specifically, the context-aware selection policy reduces the required queries by up to 88%, in contrast to the random selection method. While the number of necessary labels remains consistent for the two DQN agents throughout the analysis (as expected), the context-aware agent manages to reduce the required queries by up to 32%. These findings, derived from a subject with a higher number of labels, exhibit similar trends when extended to data from other subjects.

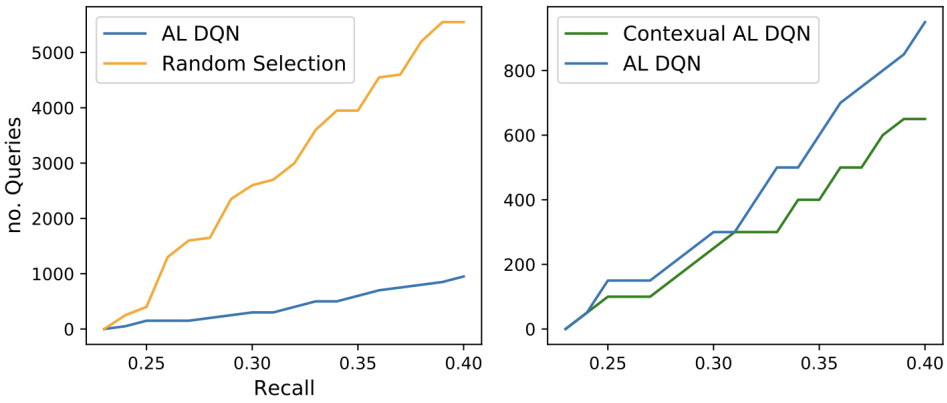


Figure 2.4: Number of queries needed to reach a certain performance level during personalization.

Personalization across Selection Policies

We examined the effectiveness of two agents and a random selection policy in achieving the primary objective of personalizing the classifier, and the findings are illustrated in Figure 2.5. The number of instances selected for querying remained consistent across each step for all selection methods. However, a subset of these selected instances stayed unlabeled, resulting in different quantities of instances available for personalization depending on the selection policies. Aside from this, the selected instances had varying impacts on personalization under different policies, comparing random selection to the other two methods.

From a single subject, we obtained a total of 12,700 instances, with 922 labeled. We reserved 230 labeled instances (from the end of the sequence) as test data from one subject, leaving the remainder for training (25% - 75%). The process started without any personalization, gradually progressing through partially labeled subjective data. A subset of this data was chosen for querying, and a part of the selected data was labeled, with the labeled data being utilized for personalization. At each step, the context-aware agent selected fewer instances than the non-context-aware agent. To ensure a fair comparison, we randomly down-sampled the number of queries from the larger group. Additionally, for random selection, we randomly

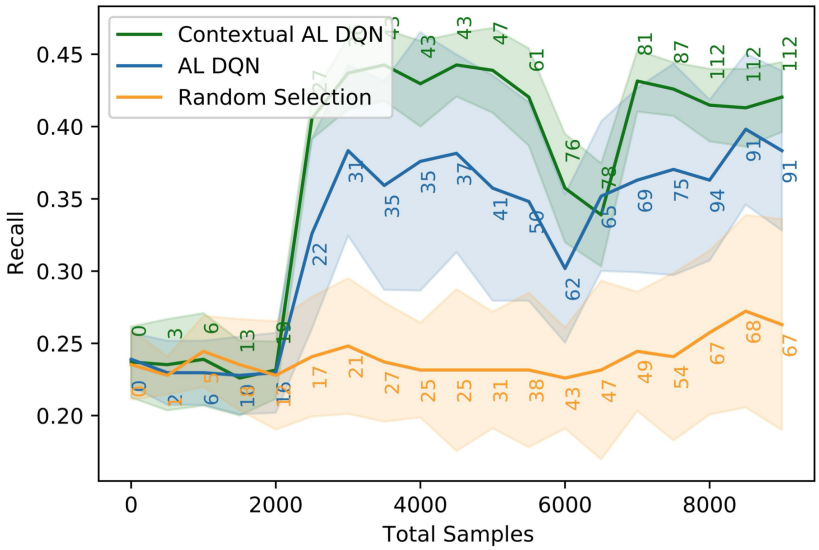


Figure 2.5: Personalization Recall in Previous Work

picked a number of partially labeled samples equivalent to the number of queries from the agents. To mitigate the impact of random selection, at each step, we selected instances and personalized the models N = 100 times. Figure 2.5 displays the mean and standard deviation of recall (True Positive Rate) for the stressed class on test data for the three selection methods.

Qualitative Summary

Instances that are selected by the random agent do not improve the performance significantly since they include samples from the entire region of the input space of the classifier, including samples whose class is 'trivial' to be extracted. Instances that are selected by a non-contextual active learning method (blue curve) increase the performance. However, with an equal number of queries, the best result is achieved when the agent is context aware (green curve), since it results in a higher number of impactful instances which also have a higher chance to receive the label.

2.5.2 Online Study: Evaluation and Results

We now present the evaluation results of the online context-aware active learning method. This complements the findings from the offline study and provides further insight into real-time stress detection performance.

In order to conduct a comprehensive comparison between our online context-aware active learning method for stress detection and previously offline variant, we have deliberately employed identical classification algorithms for both studies.

Three distinct classification techniques have been used: Support Vector Machines (SVM) [61], Random Forest [14], and XGBoost [24]. SVM finds a hyperplane in high-dimensional space to separate data classes. Random Forest uses multiple decision tree classifiers on dataset subsets, improving predictive accuracy while avoiding overfitting. Additionally, XGBoost is employed, providing an effective gradient-boosted trees implementation.

The utilization of these diverse classification techniques enables a comprehensive and robust evaluation of our proposed stress detection algorithm.

Our stress detection models are classified into two categories: single-modal and multi-modal algorithms. Within the single-modal algorithm, solely the PPG signal is employed for constructing the stress detection models. On the other hand, the multi-modal algorithm utilizes both the PPG signal and contextual information (AWARE data) in the development of the proposed models.

Experiment Detail

We use 5-fold stratified cross-validation to ensure a fair and robust evaluation across participants and to mitigate overfitting under class imbalance. K-fold cross-validation splits the

data into multiple subsets for training and testing. It helps assess model robustness across data partitions and makes fuller use of the available data. Averaging results across folds provides a stable basis for model comparison and hyperparameter selection.

Evaluation Metrics

To evaluate our stress monitoring system, we use three key metrics: F1-score, precision, and recall. The F1-score assesses binary categorization test accuracy, calculated from precision and recall, where precision measures correctly identified "true positive" results and recall identifies all "true positive" results. F1-score is a weighted average of precision and recall, important for binary classification tests.

Classification Performance Results

Table 2.4 presents a comprehensive performance analysis of our novel stress detection algorithm, incorporating an online context-aware active learning approach, compared with the offline variant.

The online cohort has higher values than the offline cohort across the reported metrics. For Random Forest, F1 increases by 0.11, from 0.21 to 0.32. Because the cohorts were collected during different periods and include different participants, this comparison does not isolate the causal effect of online triggering; it provides evidence that motivates further controlled evaluation.

This outcome also underscores the considerable advantage of incorporating contextual awareness into our model, resulting in significant enhancements across various classification metrics and reaffirming the pivotal role of context in stress detection tasks.

Table 2.4: Classification performance results (mean ± SD) across participants using 5-fold cross-validation.

Active learning method	**Data**	**Model**	**F1 (mean ± SD)**	**Precision**	**Recall**
Offline Context-Aware	PPG	Random Forest	0.21 ± 0.02	0.27	0.17
Offline Context-Aware	PPG	XGBoost	0.31 ± 0.03	0.35	0.27
Offline Context-Aware	PPG	SVM	0.41 ± 0.03	0.32	0.58
Online Context-Aware	PPG	Random Forest	0.32 ± 0.03	0.43	0.25
Online Context-Aware	PPG	XGBoost	0.39 ± 0.03	0.43	0.35
Online Context-Aware	PPG	SVM	0.50 ± 0.03	0.41	0.64
Online Context-Aware	**PPG + Context**	**Random Forest**	**0.36 ± 0.02**	**0.49**	**0.28**
Online Context-Aware	**PPG + Context**	**XGBoost**	**0.40 ± 0.02**	**0.45**	**0.36**
Online Context-Aware	**PPG + Context**	**SVM**	**0.52 ± 0.03**	**0.45**	0.61

2.5.3 Personalization Performance

Leveraging the unique physiological and behavioral variations in individuals can significantly enhance the efficacy of generic models. Inspired by the potential advantages of individualized prediction models, we hypothesize that personalizing reinforcement learning models might similarly elevate data quality. To validate this premise, we initially trained a generalized representation model using the aggregated training data from all users. Subsequently, we fine-tuned this model for each user individually, aiming to discern potential enhancements in prediction accuracy. Our evaluation centered on contrasting these generalized and personalized models to elucidate the tangible benefits of our individual-based personalization strategy in data collection.

Experiment Detail

To evaluate the effect of personalization in our data-collection mechanism, we implemented a temporal train-test splitting strategy. The dataset comprises data from multiple users. We adopted a leave-one-subject-out cross-validation scheme in which the held-out user's data are divided temporally into two parts. The initial half is used for model personalization, while the latter half is reserved for testing.

Two distinct models were constructed for comparative assessment:

- **Plain Model**: This model is trained using the entire dataset except for the data of the user currently under consideration. For testing and evaluation, the latter half of this user's data is employed.
- **Personalized Model**: This model, on the other hand, is trained using the complete dataset (excluding the data of the current user) combined with the initial half of the current user's data. Again, the latter half of the user's data is utilized for testing.

Comparing these two models helps us gauge the effectiveness of our personalization strategy. By contrasting their performance, we can see how incorporating user-specific data for training improves accuracy significantly compared to using a generic global dataset.

Personalization Performance Results

Table 2.5 and Figure 2.6 compare the personalized and non-personalized stress-detection models. The ROC curves show the relationship between the true-positive and false-positive rates across decision thresholds, while the area under the curve (AUC) summarizes discrimination performance. Personalization improves the AUC for all three classifiers; the largest increase is 0.10 for XGBoost, from 0.47 to 0.57.

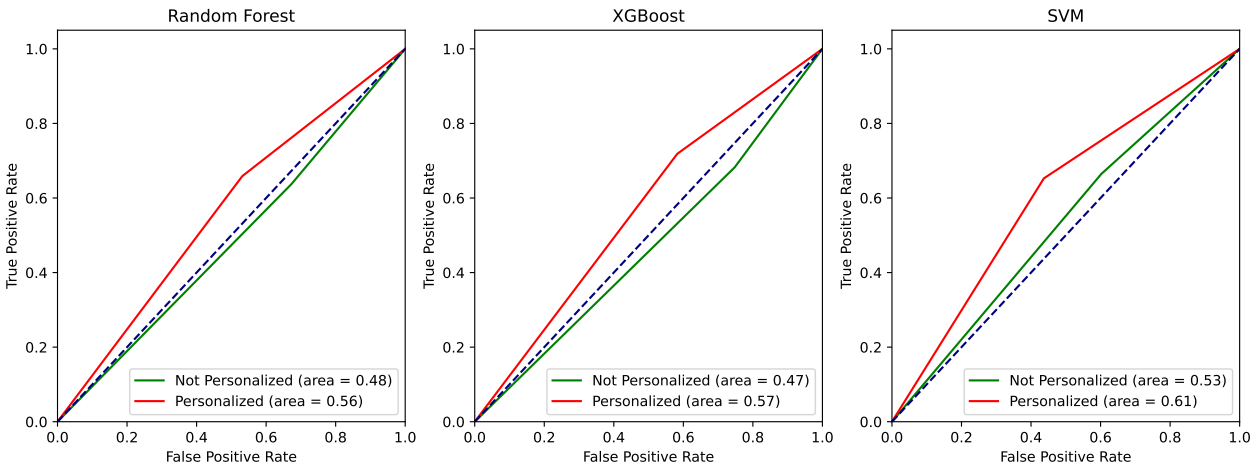


Figure 2.6: Personalization ROC Curves

Table 2.5: Personalization Results

Training method	**Model**	**F1**	**Precision**	**Recall**
Not Personalized	Random Forest	0.60	0.55	0.64
Not Personalized	XGBoost	0.60	0.54	0.68
Not Personalized	SVM	0.62	0.59	0.66
Personalized	Random Forest	**0.64**	**0.61**	**0.66**
Personalized	XGBoost	**0.66**	**0.61**	**0.71**
Personalized	SVM	**0.65**	**0.66**	0.65

2.5.4 Comparative Evaluation with Baseline Models

To evaluate the effectiveness of our proposed context-aware RL-based stress-monitoring framework, we conducted comparative experiments with traditional machine-learning and deep-learning baselines. These evaluations were performed separately for the offline and online datasets using binary stress labels.

We implemented three classical models—Random Forest, XGBoost, and SVM—which have been extensively used in prior physiological stress detection studies [55]. To explore the contribution of feature selection, each model was evaluated using both the full set of features and a subset of HRV-only features (e.g., RMSSD, SDNN, pNN20, pNN50, SD1, SD2, SD1/SD2), which are standard metrics of autonomic nervous system activity [124, 105].

Table 2.6: Performance comparison for binary stress detection (Offline setting) using classical and deep learning models. Only full-feature models are shown.

Model	**F1**	**Accuracy**	**Precision**	**Recall**
Random Forest (All)	0.53	0.63	0.57	0.50
MLP (All)	0.54	0.64	0.46	0.66
CNN (All)	0.45	0.61	0.42	0.48

Table 2.7: Performance comparison for binary stress detection (Online setting) using classical and deep learning models. All-feature models outperform HRV-only models.

Model	**F1**	**Accuracy**	**Precision**	**Recall**
Random Forest (All)	0.64	0.70	0.67	0.61
Random Forest (HRV)	0.43	0.58	0.40	0.47
MLP (All)	0.66	0.72	0.63	0.68
MLP (HRV)	0.47	0.60	0.42	0.53
CNN (All)	0.61	0.67	0.59	0.63
CNN (HRV)	0.46	0.55	0.39	0.54

Additionally, we implemented two neural network architectures to capture nonlinear patterns: a multilayer perceptron (MLP) with two hidden layers and ReLU activations, and a one-dimensional convolutional neural network (CNN) with a convolutional layer, max pooling, and a dense output. These have been applied in prior wearable emotion and stress recognition research [109, 116].

All models were evaluated using 5-fold stratified cross-validation. We addressed class imbalance using SMOTE and report results in terms of F1 score, accuracy, precision, and recall. Tables 2.6 and 2.7 present the performance of all baselines for the offline and online datasets, respectively.

2.5.5 Ablation Study on System Components

To better understand the contribution of each system component, we perform an ablation study by incrementally adding modules to a base configuration and reporting the resulting F1, precision, and recall, as shown in Table 2.8. For consistency, all results are based on a

Table 2.8: Incremental evaluation of measured online stress-monitoring configurations using a Random Forest classifier. The final row reports the complete system and does not isolate the effects of RL triggering and personalization.

System Configuration	**F1**	**Precision**	**Recall**
PPG-only + Offline Inference	0.21	0.27	0.17
+ Context Features	0.32	0.43	0.25
+ Online Inference	0.36	0.49	0.28
+ Random EMA Triggering	0.47	0.52	0.39
Full System (RL + Personalization)	**0.64**	**0.61**	**0.66**

Random Forest classifier. We start with a baseline that uses only PPG features and performs offline inference with fixed query timing. This minimal configuration yields an F1 score of 0.21, indicating limited performance due to the lack of contextual signals and real-time adaptation.

Adding contextual features such as screen activity, message count, and time of day improves the F1 score to 0.32. Transitioning to an online inference setup increases the F1 score to 0.36. A random triggering policy reaches an F1 score of 0.47. The complete system, which combines RL-based triggering and personalization, achieves an F1 score of 0.64, precision of 0.61, and recall of 0.66. Because the experiment did not evaluate RL-based triggering without personalization, the separate causal contributions of these two components cannot be estimated from these results.

Overall, the measured configurations show cumulative gains as context integration and online inference are added. The comparison supports the complete system but does not by itself isolate every component's contribution.

2.5.6 Energy and Transmission Overhead Analysis

We estimate the sensing and transmission overhead of a hypothetical high-frequency baseline that collects a 2-minute PPG window every 5 minutes. This baseline is distinct from the deployed study schedule, which collects a 2-minute window every 15 minutes. Because hardware-level energy measurements were not performed, the estimates use published specifications for representative commercial-grade sensors and communication protocols.

Sensing Energy. We model PPG sensing energy consumption using the MAX30101 optical sensor, which draws approximately 5.5 mW during active operation and 3.5 μW in standby mode [97, 78]. Under the hypothetical 5-minute baseline and a 24-hour window, the estimated daily sensing energy is 190.26 J.

Table 2.9: Estimated Daily Energy Consumption for Baseline Method

Component	**Baseline (Continuous Sensing)**
PPG Sensing Energy (5.5 mW active mode)	190.26 J
BLE Transmission Energy (1 Mbps BLE 5)	415.38 mJ

Transmission Energy. Data transmission is modeled based on Bluetooth Low Energy (BLE 5.0) using parameters from Silicon Labs documentation [76]. Assuming a data rate of 1 Mbps, a power draw of 18.78 mW during active transmission, and a sampling rate of 20 Hz with 4-byte float values, the estimated total transmission energy for the baseline method is 415.38 mJ per day.

We do not report measured energy consumption for the deployed method because direct runtime measurements were not performed. Its lower sampling frequency implies fewer sensing and transmission events than the hypothetical baseline, but the resulting energy savings should be verified through device-level measurement.

2.6 Discussion

This study presents a context-aware RL framework for optimizing the timing of EMAs in stress monitoring applications. While our results demonstrate improvements in user engagement, sensing efficiency, and classification performance, there are several limitations and broader implications that warrant discussion.

2.6.1 System Limitations and Real-World Generalizability

During online deployment, the core stress detection model and reinforcement learning policy parameters remain fixed for a given period and are updated only through scheduled retraining using newly collected labeled data. Continuous online parameter tuning was not performed, as the study prioritizes deployment stability and participant safety in a long-term human-subject setting. Our implementation is currently limited to a controlled environment using a Samsung Galaxy Active 2 smartwatch and a paired Android smartphone. The sensing frequency, communication interval, and available contextual data are specific to this device setup. While this serves as a valuable proof of concept, real-world deployments will require adaptation to heterogeneous hardware, including wearables with different sampling rates, operating systems, and communication protocols. We discuss these limitations explicitly and outline potential solutions, such as device-specific policy fine-tuning and adaptive feature scaling, for future deployments.

Additionally, our participant pool primarily consisted of undergraduate and graduate students. While this cohort provides a meaningful testbed for feasibility, it limits the generalizability of our findings. Stress patterns and engagement behaviors may vary significantly across demographics, including older adults, professionals, and clinical populations. To improve ecological validity, future work will expand data collection to include more diverse

cohorts and stress-inducing environments beyond academic life. Deploying the system in an online setting introduced practical challenges not present in offline evaluation, including real-time data transmission, intermittent connectivity, delayed or missing user responses, sensor noise, error handling, and the need to balance EMA frequency with sustained participant engagement. Evaluating the proposed approach under these realistic constraints is a primary motivation of the online study.

2.6.2 Privacy Considerations

Wearable health systems process sensitive physiological and contextual data, making privacy protection essential. In the deployed architecture, raw sensor and contextual data are transmitted from the paired devices to ZotCare cloud services for processing and storage. The system therefore requires secure transmission, access control, and storage practices consistent with the approved study protocol. Future implementations could reduce exposure through on-device feature extraction or inference and could incorporate federated learning and differential privacy.

Building upon this work, our recent study [141] demonstrates a differentially private federated transfer learning framework for stress detection, which provides robust privacy guarantees while maintaining model performance. We envision incorporating similar techniques into future versions of our system to address data privacy and security concerns at scale.

2.6.3 Reward Function Design and Policy Interpretability

The reinforcement learning policy that drives EMA triggering was trained using a reward function that balances sensing cost with anticipated label utility. The coefficients in this function were derived through empirical hyperparameter tuning. The mathematical formulation

reports these values to support transparency and reproducibility. In future work, automated reward-tuning methods, such as meta-optimization or inverse reinforcement learning, could reduce heuristic bias and enhance policy interpretability.

2.6.4 User-Centered Evaluation and Engagement

This study evaluates engagement indirectly through query efficiency and the availability of usable EMA labels. It does not include qualitative assessments of participant experience, and user perceptions of prompt timing, burden, and trust remain unmeasured. Future studies should incorporate diary studies, interviews, and usability metrics to evaluate these dimensions directly and guide refinement of the triggering strategy and user-facing interfaces.

2.6.5 Stress Modeling: Binary vs. Continuous

Our experiments adopt binary stress classification, in line with common practices in stress-monitoring literature. However, stress is inherently multidimensional and continuous, and binarization discards distinctions among adjacent self-reported levels. Evaluating multi-class, ordinal, or continuous stress models remains an important direction for future work, particularly for personalized wellness applications.

2.6.6 Statistical Reliability and Class Imbalance

All reported performance metrics are based on stratified 5-fold cross-validation, and standard deviations of F1 scores across participants reflect variability. This reporting improves the robustness and interpretability of the evaluation.

Class imbalance is expected in real world health data because participants experience fewer

high stress episodes than low stress episodes. We use SMOTE only within each training fold; validation and test folds retain their original distributions. This prevents synthetic samples from altering evaluation prevalence while allowing the classifier to learn from a less imbalanced training set. The RL-based decision-making framework further adapts EMA frequency based on expected label utility.

2.6.7 Model Comparisons and Baselines

To address concerns about feature selection and baseline comparisons, we expanded our experiments to include both traditional HRV-based features and representative machine-learning and deep-learning baselines, including Random Forest, XGBoost, SVM, MLP, and CNN models. These comparisons show that the full-feature models outperform their HRV-only counterparts in the online setting. The incremental system comparison further quantifies measured gains from context features and online inference, while the separate effects of RL triggering and personalization remain a limitation.

2.6.8 Computational Overhead and Energy Usage

We estimated the daily energy consumption of a hypothetical high frequency PPG sensing and Bluetooth transmission baseline using specifications from commercial sensors. These estimates contextualize baseline system costs, but we could not profile the complete latency and energy use of the adaptive framework because device-level instrumentation was unavailable. Future work should conduct on-device profiling to evaluate latency, runtime load, and real-time responsiveness.

2.6.9 Future Directions

Our findings provide a promising starting point for intelligent, context-aware EMA scheduling in real-world stress monitoring. Looking ahead, we aim to (i) deploy the system across larger and more diverse populations, (ii) incorporate longitudinal user feedback, (iii) model stress continuously or ordinally, (iv) implement differential privacy and federated learning, and (v) optimize the RL reward function through automated techniques. These extensions will help improve the scalability, interpretability, and fairness of next-generation wearable mental health systems.

2.7 Conclusions

In conclusion, this work introduced a novel contextual variant of active learning, leveraging Deep Q-Learning to incorporate individual contextual information into the decision-making process [126, 125]. In the initial phase, the implementation of a context-aware active reinforcement learning algorithm in an offline setting showcased its efficacy, resulting in a significant reduction of up to 88% in required EMAs compared to random selection and up to 32% compared to traditional active learning methods. Additionally, stress detection performance exhibited notable improvements, with up to a 21% enhancement compared to random selection and up to 8% compared to traditional active learning.

Moving to the second phase, our online implementation used active learning to initiate EMAs based on real-time contextual information. Comparisons across the offline and online study cohorts showed an increase of up to 11% in stress-detection performance for the online cohort under the reported evaluation. Incorporating contextual features improved results by an additional 4%, while personalization further improved model performance.

This study not only contributes a valuable advancement in stress detection methodologies

but also underscores the pivotal role of context-awareness and online implementation in achieving superior results. The demonstrated reductions in participant burden and improvements in label accuracy signify the potential practical impact of this research in real-world applications. Future directions may explore additional personalization techniques and extend the application of context-aware active learning to diverse domains, fostering continued advancements in intelligent and user-centric systems.

Data Availability

The authors are preparing de-identified versions of the datasets for public release in accordance with institutional and IRB data-sharing requirements. Repository metadata and a persistent link will be added when the release is complete.

Chapter 3

Differential Private Federated Transfer Learning for Mental Health Monitoring in Everyday Settings: A Case Study on Stress Detection

This chapter is based on material published in the proceedings of the 2024 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC).

3.1 Introduction

Mental health concerns impact a diverse demographic cross-section of the population [99]. The prolonged effects of these mental health conditions can have detrimental impacts on physical health [27, 146], psychological states, and overall quality of life [158, 28, 157].

In response, the field of mental health monitoring has increasingly turned to data-driven methodologies [15, 3]. These methods, utilizing advanced analytics and machine learning algorithms, provide precise, real-time assessments of various mental health states [51]. However, the ascendancy of data-driven techniques in mental health monitoring brings forth pressing concerns regarding data privacy [67]. The sensitive nature of the data involved demands methodologies that are not just accurate and efficient in monitoring mental health but also stringently protective of individual privacy [140, 48, 2].

Recent studies in mental health monitoring (e.g., mood, loneliness, depression, stress) increasingly adopt federated learning to enhance privacy [143, 1, 26, 122, 100, 18, 147], showing promise in using local data for model training. For example, [122] developed a framework for depression detection using smartphone data (location, accelerometer, call logs) for training on devices, although vulnerable to privacy breaches like membership inference attacks. [100] and [18, 112] similarly leveraged federated learning for loneliness and stress detection, utilizing sensor and PPG signal data from smartphones and wearables, respectively, to update models locally, thus preserving personal data privacy.

Although existing approaches have shown success in enhancing privacy in certain domains, they still exhibit vulnerabilities to a range of cyber threats, notably backdoor and inference attacks. The absence of more sophisticated privacy-preserving mechanisms, such as those offered by differential privacy, represents a gap in the current approaches for mental health monitoring. Furthermore, a significant hurdle in the practical deployment of federated learning in mental health studies stems from the reliance on limited data sources [65, 30]. When models are trained locally in such studies, there's often a marked inconsistency in the diversity and volume of data, as noted by Dixit et al. [34]. This variability can lead to inadequate training of decentralized models, especially when real-world data lacks the comprehensiveness or balance needed for effective training [29]. Therefore, resolving this issue is imperative for the effective application of federated learning in practical scenarios.

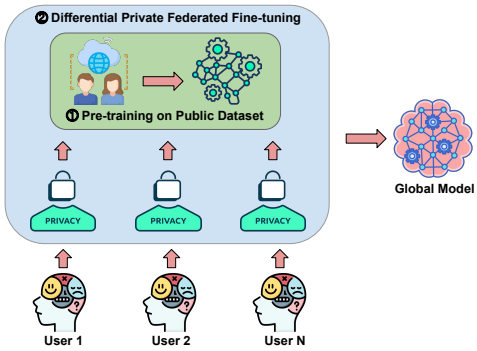


Figure 3.1: Overview of the proposed framework.

In this study, we propose a differential private federated transfer learning framework for mental health monitoring. We incorporate a differential privacy mechanism into federated learning by introducing noise into the combined updates prior to their transmission back to the central server. Furthermore, we utilize transfer learning to tackle data insufficiency and imbalance in stress detection as a case study. Specifically, we commence by pre-training a universal model on an extensive, non-sensitive dataset, subsequently refining it with individual user data on-device. We evaluate the proposed framework via a case study on stress detection on a dataset from a longitudinal study. The dataset includes physiological and contextual data gathered from Samsung Galaxy Active 2 Watches and smartphones from 54 individuals. The label stress levels of these individuals were collected through multiple daily ecological momentary assessments (EMAs) conducted via their smartphones.

3.2 Methodology

Our methodology addresses data scarcity and privacy challenges in mental health monitoring by integrating transfer learning with federated learning, enhanced by differential privacy techniques. We initially pre-train a foundational model on a comprehensive public dataset to capture broad health trends. This model is then fine-tuned on sparse, user-specific data in a distributed manner, improving accuracy with limited data. Differential privacy is applied

during weight updates to protect individual data privacy. This combined approach effectively balances privacy concerns with the need for precise health monitoring. Our methodology is depicted in Figure 3.1.

3.2.1 Differential Private Federated Transfer Learning Framework for Mental Health Monitoring

The learning process consists of three stages, which are explained in depth in the subsequent sections. First, a universal model is pretrained on a large public dataset. Second, each client fine-tunes a copy of that model using private local data, clips the resulting gradients, applies Laplacian noise, and returns a protected update to the server. Third, the server aggregates the protected client updates to produce the next global model.

Pre-training

In real-world scenarios, gathering extensive, high-quality data for mental health studies is often fraught with difficulties. Participants in such studies may not consistently engage, leading to gaps or irregularities in the data. Additionally, the collected data often contains artifacts and noise, further complicating its utility and quality. These factors combine to create a scenario where the available data is scarce, inconsistent, and noisy.

We initiate our approach by pre-training our model $\mathcal{M}$ on a comprehensive publicly available dataset $\mathcal{D}$pt. This dataset, rich in both quantity and quality, comprises N pairs of well-curated features and labels:

$$D_{pre-train} = (x_1, y_1), (x_2, y_2), ..., (x_N, y_N). \tag{3.1}$$

The pre-training phase utilizes the binary cross-entropy loss, and is aimed at learning generalized mental health patterns:

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \cdot \log(\mathcal{M}(x_i)) + (1-y_i)\cdot \log(1-\mathcal{M}(x_i))\right]. \tag{3.2}$$

Differential Private Federated Fine-tuning

To tackle the critical privacy concerns in mental health monitoring, our approach incorporates differential privacy (DP) within the federated learning paradigm, specifically through the Federated Averaging (FedAvg) algorithm [82] enhanced with Laplacian noise. This implementation is key to defending against sophisticated privacy attacks against traditional federated learning frameworks, including model inversion [50], membership inference [87], and backdoor [10], which pose risks to user data confidentiality.

Federated Averaging FedAvg is a widely-used algorithm in federated learning that involves averaging model updates (weights) from multiple clients. In our implementation, each client represents a participant in the mental health monitoring study, first trains a local model using their private data. The local models are then aggregated to update the global model. This process is formalized as:

$$\mathcal{M}_{\text{global}} = \frac{1}{K}\sum_{k=1}^{K}\mathcal{M}_k, \tag{3.3}$$

where $\mathcal{M}_{\text{global}}$ represents the global model, $\mathcal{M}_k$ is the model trained on the k-th client, and K is the total number of clients.

Laplacian Noise for Differential Privacy To ensure differential privacy, we add Laplacian noise to the model updates during aggregation. The noise is calibrated to the sensitivity of the model's gradients and the privacy budget ϵ. For a model parameter θ, the update with Laplacian noise is:

$$\theta_{\text{updated}} = \theta + \text{Lap}\left(\frac{\Delta f}{\epsilon}\right), \tag{3.4}$$

where Δf denotes the sensitivity of the function (gradient norm), and $\text{Lap}(\cdot)$ represents the Laplacian noise.

Fine-tuning Following the pre-training phase, the model $\mathcal{M}$ utilize private, user-specific datasets, represented as $D_{fine-tune} = \{(x'_j, y'_j)\}_{j=1}^{M}$ to fine-tune. This process occurs within a federated learning framework and is essential for adapting the generalized model to the specific mental health profiles and solve the data scarcity issue of the private dataset. During fine-tuning, the model parameters are adjusted according to each user's data. The fine-tuning employs the same loss function, as Eq. 3.2, used in the pre-training phase, ensuring consistency in the optimization approach across both stages of model development.

3.2.2 Case Study on Stress Detection

The efficacy of our framework is showcased through a case study in stress detection, leveraging data from an extensive longitudinal study [126]. This study engaged distinct groups of participants across its two phases, with 30 individuals in the initial phase and 24 in the subsequent phase, including both undergraduate and graduate students. Data collection occurred over two periods: from June 2020 to June 2021 and from March 2022 to May 2023, resulting in 109,586 and 23,012 refined samples for each period, respectively. The dataset from the first phase, $D_{pre-train}$, was critical for the pre-training process, and the dataset

from the second phase, $D_{fine-tune}$, was essential for fine-tuning. Conducted using our ZotCare health monitoring system [115], this strategy leveraged the comprehensive dataset from the first phase for initial model training, while the more constrained dataset from the second phase was used for precise model enhancement, effectively tackling the challenges posed by data scarcity.

Dataset

The collected dataset comprises raw PPG and motion data, along with partial annotations for stress levels, emotions, and physical activity through EMA at semi-random intervals. To comprehensively process the PPG signals, we employ the HeartPy library [128], extracting 12 specific features from both electrical activations and pressure waveforms in the dataset. These features[1] are heart rate (HR) and heart rate variability (HRV) measures. Stress levels equal to or less than 2 are categorized as 'unstressed,' while higher values indicate 'stressed' individuals.

Stress Detection Model

To develop our stress detection model, we began with a pre-processing stage to refine bio-signals from wearable devices. Initial data cleaning eliminated erroneous readings by using motion data to filter out noise and artifacts, utilizing a bandpass Butterworth filter specific to PPG signal frequencies. This was followed by a moving average technique to smooth the data, minimizing motion-related distortions. We then applied min-max normalization, scaling feature values uniformly between 0 and 1, to reduce individual data variations and enhance bio-signal interpretability.

To model stress levels, we opted for a Multilayer Perceptron (MLP). This MLP is structured

[1]BPM, IBI, SDNN, SDSD, RMSSD, pNN20, pNN50, MAD, SD1, SD2, S, SD1/SD2, and BR1.

with a three-layer design, including hidden layers that consist of 128 and 32 neurons, respectively. It processes an input comprising the 12 extracted features related to heart rate (HR) and heart rate variability (HRV). The output of the MLP is a binary classification, differentiating between states of stress and no stress.

3.3 Evaluation

This section is dedicated to the evaluation of our proposed framework. We establish a baseline using the pre-trained model to highlight the benefits of incorporating transfer learning. Additionally, we explore the delicate balance between privacy preservation and model accuracy by adjusting the privacy budget parameter, ϵ.

Implementation Details

We train the proposed stress detection model by the Adam optimizer (learning rate=0.001) and cross-entropy loss function. A dropout strategy (rate=0.5) is implemented after the initial hidden layer to reduce overfitting. Differential privacy via Laplacian noise addition to the gradients was introduced, with an ϵ value of 1, aiming for a balance between privacy protection and model accuracy.

We pre-train on $D_{pre-train}$ with 3271 labeled samples for 50 epochs to capture a wide array of stress signals. Subsequently, we applied FedAvg to fine-tune the model on $D_{fine-tune}$, consisting of 1220 labeled samples, over 30 epochs for personalized user data adaptation.

The dataset $D_{fine-tune}$ was divided in a manner that allocated the last 30% of each user's chronologically ordered data to the test set, ensuring the model was trained on past data and evaluated on future data. Client partitioning for federated learning was conducted based

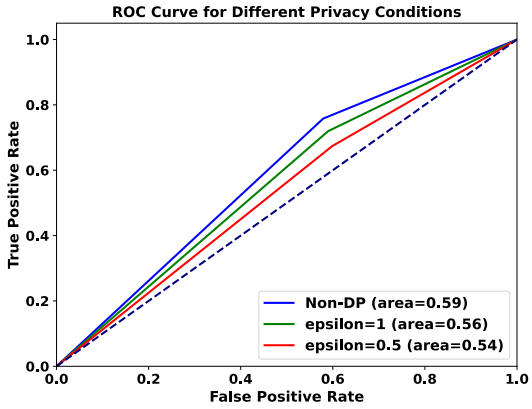


Figure 3.2: Overview of the proposed approach.

on individual users, reflecting a real-world scenario where fine-tuning occurs on each user's client device in a personalized manner.

Results

The results of our proposed framework for stress detection are shown in Table 3.1. The Plain model, trained solely on the original dataset $D_{fine-tune}$ without incorporating transfer learning techniques, yielded an accuracy of 0.43, an F1 Score of 0.39, a recall of 0.54, and a precision of 0.31. Conversely, training the model on the public, more extensive dataset $D_{pre-train}$ resulted in improved metrics: accuracy rose to 0.51, F1 Score to 0.44, recall to 0.25, and precision to 0.36. Notably, the integration of transfer learning within our differential privacy federated learning framework enhances performance, achieving an accuracy of 0.53, an F1 Score of 0.52, a recall of 0.75, and a precision of 0.40.

Table 3.1: Stress Detection Performance of Different Models

Model	**Accuracy**	**F1 Score**	**Recall**	**Precision**
Plain	0.43	0.39	0.54	0.31
Pre-trained	0.51	0.44	0.58	0.36
Fine-tuned	**0.53**	**0.52**	**0.75**	**0.40**

Privacy Budget Analysis

In the final stage of our analysis, we conducted comparisons of our outcomes with two reference benchmarks: one involving the identical neural network framework without the implementation of federated learning or differential privacy, and the other utilizing the same neural network structure with federated learning but absent differential privacy. This comparative evaluation was performed across two scenarios lacking differential privacy, with ϵ values set at 0.5 and 1, respectively. It's important to note that a lower ϵ value signifies enhanced privacy safeguards. As depicted in Figure 3.2, integrating differential privacy within a federated learning context did not significantly alter the metrics of the receiver operating curve (ROC). This observation suggests that it is feasible to strike an effective equilibrium between preserving model efficacy and upholding privacy standards.

3.4 Discussion

In this study, we introduce a differential privacy federated transfer learning framework and evaluate it on stress detection. By applying transfer learning, our approach effectively leverages the strengths of both public and private data sources, enhancing the model's performance in real-world scenarios. Especially noteworthy is the improvement in recall indicates that our framework is effective in reducing potential overlooking true stress cases. This advance is crucial, minimizing missed detections of stress and thereby mitigating their potential adverse effects on individual well-being and public health. Additionally, the implementation of differential privacy within the framework ensures the protection of sensitive health data, which is a critical concern in the field of mental health. By infusing noise into the model updates, our framework mitigates risks associated with several cyber attacks [50, 87, 10].

This study's limitation is its focus on stress detection, which may not encompass other

mental health conditions or capture long-term trends effectively due to its design for short-term analysis. Future efforts could aim to broaden the framework's applicability to various mental health states and for different cohorts. This could improve its predictive power for both immediate and extended periods that were shown feasible in prior research [145].

3.5 Conclusion

This study proposed a differential private federated transfer learning framework addressing the challenges of data scarcity and privacy protection in mental health monitoring. The proposed framework exhibited strong performance in stress detection while effectively preserving data privacy. Specifically, our framework achieved a substantial 10% improvement in accuracy and a noteworthy 21% enhancement in recall. These results highlighted the potential of our approach to provide more effective and privacy-conscious mental health monitoring, addressing the pressing need for efficient solutions in this domain.

Chapter 4

ECG Unveiled: Analysis of Client Re-identification Risks in Real-World ECG Datasets

This chapter is based on material published in the proceedings of the 2024 IEEE 20th International Conference on Body Sensor Networks (BSN).

4.1 Introduction

The digitization of health records and the proliferation of wearable biosensors have revolutionized e-health [3], enabling continuous monitoring and real-time analysis of physiological signals [3, 7]. Among these, Electrocardiogram (ECG) signals capture the heart's electrical activity through distinct PQRST complexes, providing vital insights into heart health and enabling the detection of various cardiac abnormalities [46]. While ECG signals are primarily used for medical diagnosis and treatment, they have unique biometric properties that can

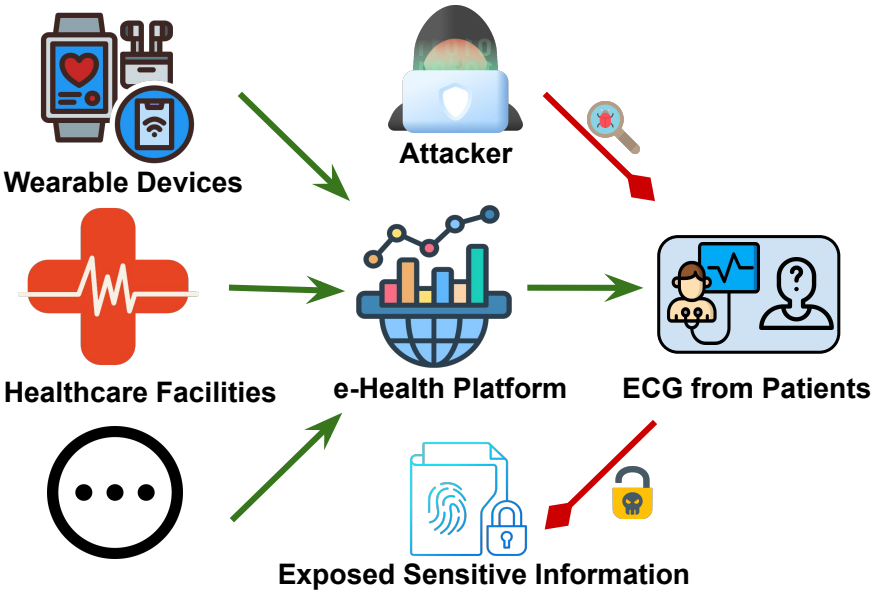


Figure 4.1: Overview of the threat model illustrating ECG data aggregation from various sources to e-health platforms, creating an attack surface for potential client re-identification.

be exploited to identify individuals [52].

As ECG data becomes more accessible through digital-health platforms and health record databases, the risk of client re-identification from public datasets significantly increases [91, 141]. Public datasets are essential for research in healthcare. These datasets can be exploited through various machine learning-based attacks, such as linkage attacks [98] and membership inference attacks [113], further exacerbating client re-identification risks (Figure 4.1). Attackers can leverage the distinct biometric features of ECG signals to re-identify clients, leading to significant privacy breaches [48] and potential misuse of personal health information. This highlights the urgent need for robust privacy-preserving mechanisms to safeguard clients' identities in publicly accessible datasets [41].

Recent studies have shown that ECG signals can be used for biometric authentication, revealing privacy risks as ECG statistical variants can disclose individual identities [52]. These vulnerabilities allow for pinpointing individuals within subgroups based on demographic information and health conditions [42, 47]. Similar privacy risks exist in other biosignal

data types, such as Photoplethysmograph (PPG) and Electroencephalogram (EEG), highlighting the need for robust privacy-preserving techniques across all biosignal health systems [140, 147]. The unique variants of ECG signals, illustrated by the PQRST peaks and key features (Figure 4.2), can be exploited for client re-identification. Existing studies often rely on homogeneous datasets and controlled conditions that fail to capture the diversity and complexity of practical applications [42]. For example, homogeneous datasets may consist of ECG recordings from a single demographic group or clinical setting, while controlled conditions involve standardized environments that do not reflect everyday variability. In contrast, real-world data is inherently diverse, covering various demographic groups, clinical conditions, and recording environments. Therefore, a comprehensive study and analysis of client re-identification risks in diverse, real-world ECG datasets is needed to better understand and mitigate these risks.

Research Gaps and Our Contributions

The primary challenges in current research include the following: (i) existing research has not adequately considered real-world threat models, limiting their applicability [52]; (ii) the reliance on deep learning models is problematic because these models extract latent space features that lack explainability and cannot be interpreted by humans [30], leaving domain experts without the necessary information to develop effective privacy-preserving mechanisms [40]; and (iii) previous works did not thoroughly investigate the factors contributing to client re-identification risks. Their methods often rely on trial experiments or observations on small samples, leading to inadequate and inefficient examination of key features [42].

In this study, we address these gaps by investigating the re-identification risks associated with ECG data using transparent machine learning models. A key aspect of our approach is the use of SHAP analysis to identify the most influential features contributing to re-

identification. By building interpretable models, we evaluate the effectiveness of ECG signals in re-identifying individuals across diverse demographic groups and clinical conditions within real-world heterogeneous datasets. This analysis provides valuable insights that can guide the development of privacy-preserving techniques and inform clinical experts.

In summary, our contributions are as follows:

- We provide a comprehensive analysis of re-identification risks using transparent machine learning models, applied to heterogeneous datasets from multiple sources, to accurately reflect real-world scenarios.
- We offer findings that highlight the need for robust anonymization techniques and privacy-preserving mechanisms, providing a foundation for future research aimed at protecting sensitive health information.
- We conduct a rigorous investigation into the factors contributing to re-identification, utilizing feature importance assessment and SHAP analysis to provide insights that are interpretable by domain experts.

4.2 Analysis of Re-identification Risks

4.2.1 Method

Our primary objective is to assess the re-identification risks associated with ECG data. To achieve this, we extract key PQRST features, which represent distinct electrical activities of the heart and are crucial for identifying individual-specific patterns [66]. Focusing on these features allows us to build interpretable models that reveal re-identification factors and provide actionable insights for clinical experts.

Feature Extraction from ECG Signals

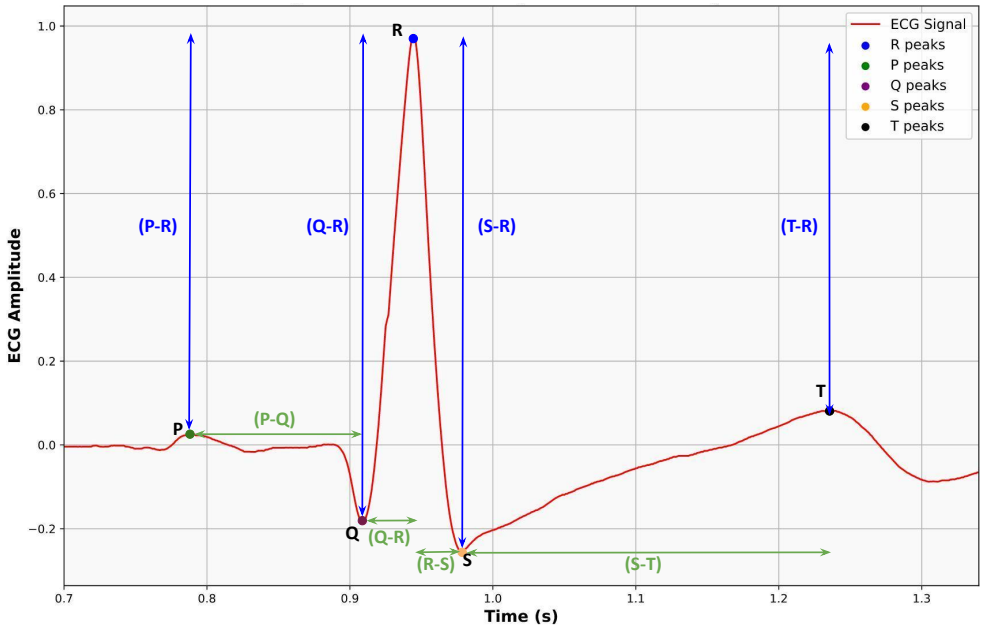


Figure 4.2: ECG signal with detected PQRST peaks and key features, highlighting its biometric properties.

We extracted key PQRST features from ECG signals using the NeuroKit2 [39] library. The process involved three main steps: ECG signal cleaning, R-peak detection, and delineation of P, Q, S, and T peaks to identify precise fiducial points. Signal cleaning was essential to address noise and baseline wander in raw ECG signals. We applied a highpass Butterworth filter to remove slow drifts and direct current (DC) bias, followed by a powerline filter to eliminate 50 Hz interference from electrical sources.

To capture distinctive ECG characteristics for identifying patterns associated with individuals, we derived several statistical features, such as mean amplitude differences and interval durations. Previous studies have shown that these statistical features effectively distinguish individual-specific cardiac patterns in the PQRST complex, making them suitable for re-identification tasks [52, 42]. Specifically, the P, Q, R, S, and T peaks represent different phases of the cardiac cycle and are critical for capturing the heart's electrical activity [66]. These peaks can be used to extract clinically significant features. We computed the Mean

Amplitude Differences between the P, Q, S, T peaks and the R-peak:

$$\text{Mean Amplitude Difference (X-R)} = \frac{1}{N}\sum_{i=1}^{N}(A_{X,i} - A_{R,i}) \tag{4.1}$$

where X represents the P, Q, S, or T peaks, A_X and A_R are the amplitudes at these peaks, and N is the number of valid peak pairs. Additionally, we calculated the Mean Intervals between the clinically significant P, Q, R, S, and T peaks to capture temporal features. These include:

$$\text{Mean Interval (X-Y)} = \frac{1}{N}\sum_{i=1}^{N}(t_{Y,i} - t_{X,i}) \tag{4.2}$$

where X and Y represent different peaks (i.e., P, Q, R, S, T), and t_X and t_Y are the times of these peaks.

As illustrated in Figure 4.2, we identified and highlighted the key PQRST features on an ECG signal. The figure shows the amplitude differences (e.g., $P-R$, $Q-R$, $S-R$, $T-R$) and intervals (e.g., $P-Q$, $Q-R$, $R-S$, $S-T$) between significant points. Unlike deep learning models, which often generate features in latent spaces that are difficult to interpret and understand [79], these explainable features capture unique physiological variations in heart activity among individuals, considering demographic, health, lifestyle, and genetic factors, which are particularly effective for re-identifying individuals [44].

Data Processing

Our experiments targeted three main tasks: binary gender re-identification, age group re-identification, and participant ID re-identification. These tasks were chosen because they are common targets in privacy leakage practices, making understanding their re-identification

Table 4.1: Summary of ECG Datasets Used in Experiments

Dataset	**Subjects**	**Age range**	**Sex (M/F)**	**Rate (Hz)**	**Health condition**
MIT-BIH Arrhythmia [85]	47	23–89	25/22	360	Arrhythmias
SHAREE [83]	139	55–72	90/49	128	Hypertension
BIDMC CHF [11]	15	22–71	11/4	250	Congestive heart failure
Brno University of Technology [89]	15	21–83	9/6	1000	General population
MIT-BIH Long-Term [85]	7	46–88	6/1	128	Long-term monitoring
Combined	**223**	**21–89**	**141/82**	**Multiple**	**Multiple**

risks crucial [45]. Binary gender and age group re-identification are essential due to their frequent use in demographic analyses and targeted services, where age-related information is often sensitive [147, 45]. Participant ID re-identification assesses the risk of uniquely identifying individuals within a dataset, highlighting the privacy implications in longitudinal data [91]. The data splitting methodologies for these tasks are summarized in Table 4.2.

Table 4.2: Data Splitting Methods for Re-identification Tasks

Target	**Train**	**Test**	**Label**
Gender	80% participants	20% participants	Binary
Age Group	80% participants	20% participants	Age group[a]
Participant ID	First 80% of each participant's data	Last 20% of each participant's data	Unique ID

Note: [a]Age groups: 21–30, 31–40, 41–50, 51–60, 61–70, and 71–89 years.

4.2.2 Experiment

Model Training To investigate re-identification risks in ECG data, we used interpretable and explainable models including *logistic regression*, *decision trees*, and *random forest*. Logistic regression offers straightforward coefficient interpretation, decision trees provide clear decision paths, and random forests offer feature importance scores [43]. SHAP works well

with these models by attributing each feature's contribution to the prediction, providing consistent and locally accurate explanations. This approach helps identify key features and provides actionable insights for clinical experts, ensuring the safe use of ECG data while maintaining privacy.

Model Evaluation and Interpretability Analysis The tuned models were evaluated using standard classification metrics, including accuracy, precision, recall, F1-score, and ROC AUC, and confusion matrices were generated to visualize performance across different classes. We conducted SHAP analysis to understand the contributions of each feature to the model's predictions. Let f be the model, x be the feature vector, and x_i be the value of the i-th feature. SHAP values $\phi_i(x)$ represent the contribution of x_i to the prediction $f(x)$:

$$f(x) = \phi_0 + \sum_{i=1}^{M} \phi_i(x) \tag{4.3}$$

where ϕ_0 is the base value (mean prediction) and M is the number of features.

Dataset In our experiments, we utilized five distinct real-world ECG datasets, each offering a diverse range of demographic and clinical conditions (refer to Tab. 4.1). For each dataset, we extracted appropriate segments (30-120 minutes) of ECG recordings to capture continuous data and reflect real-world scenarios. This diverse, multi-sourced approach enhanced the robustness of our assessment of re-identification risks associated with ECG data.

4.2.3 Results

Re-identification Risk

The models were trained and validated on data from the multi-sourced datasets, comprising 223 participants. The gender re-identification model achieved an accuracy of 0.755, the

Table 4.3: Performance Metrics for Re-identification Models

Re-identification Task	**Accuracy**	**Precision**	**F1**
Gender	0.755	0.766	0.760
Age Group	0.671	0.623	0.633
Participant ID	0.819	0.817	0.810

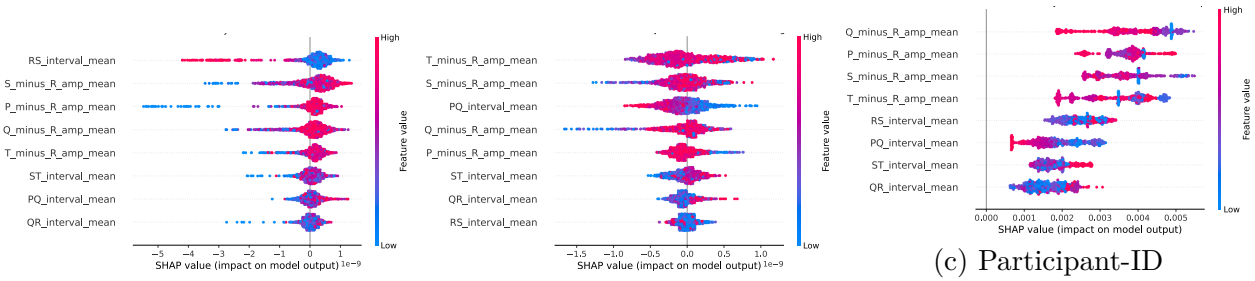
age group re-identification model achieved an accuracy of 0.671, and the participant ID re-identification model achieved an accuracy of 0.819. Detailed performance metrics are presented in Table 4.3.

High accuracy in gender and age group classification indicates that even without access to the target client's specific data during training, it is possible to infer the data owner's age range and gender using a small segment of ECG data or statistical features. This capability poses a significant threat to privacy, as adversaries can categorize individuals based on demographic attributes from minimal data. For participant ID re-identification, the model's high accuracy (0.819) suggests that an attacker can confidently match a small segment of a client's ECG data to their identity among 223 participants. This poses a severe threat to e-health systems, as depicted in Figure 4.1, facilitating unauthorized identification and potential misuse of personal health information.

These privacy breaches can lead to unauthorized access to sensitive health data, discrimination based on health status, and loss of trust in e-health systems. E-health systems have increasingly relied on machine learning enhancements in recent years to improve diagnostics, personalize treatment plans, and predict patient outcomes [141, 140]. Consequently, these systems often utilize Machine Learning as a Service (MLaaS) to manage large datasets and deploy complex models efficiently. However, unauthorized information obtained through re-identification can be combined with other attacks, such as attribute inference and membership inference attacks, further compromising privacy [113]. As highlighted by studies on linkage attacks and profiling attacks [98], these risks threaten the integrity of MLaaS systems. Therefore, robust privacy-preserving techniques are crucial to mitigate these risks

and protect individual privacy when deploying ECG data in clinical and research settings.

Re-identification Factors



(a) Gender re-identification (b) Age-group re-identification (c) Participant-ID re-identification

Figure 4.3: SHAP analysis for re-identification tasks: (a) gender, (b) age group, and (c) participant ID.

The combined analysis in Figure 4.3 reveals the factors behind re-identification risk and offers actionable insights for clinical experts. Notably, features such as S–R and P–R amplitude differences were consistently significant across tasks, indicating their strong influence on re-identification.

For gender re-identification (Figure 4.3a), the R-S interval and S-R amplitude difference were prominent. Clinically, the R-S interval, representing the time between the peak of the R wave and the end of the S wave, varies significantly due to gender differences in cardiac structure and function [66]. Larger S-R amplitude difference can indicate variations in ventricular depolarization, which often differ between men and women due to anatomical and physiological differences in the heart [44]. Additionally, differences in P-R amplitude difference reflect anatomical differences, which can be influenced by gender-specific factors such as hormonal effects on the autonomic nervous system.

In age group re-identification (Figure 4.3b), the T-R amplitude and P-Q interval played crucial roles. Age-related changes in the cardiovascular system, such as increased arterial stiffness and altered atrial conduction, affect these intervals [44]. The variability in T-R am-

plitudes among different age groups reflects these physiological changes. The P-Q interval, which measures atrial to ventricular conduction time, also varies with age due to structural and functional changes in the heart's conduction system. For participant ID re-identification (Figure 4.3c), Q-R amplitude difference and P-R amplitude difference were particularly impactful. The Q-R amplitude represents the voltage difference between the Q and R waves, which can vary significantly among individuals due to differences in myocardial mass and conduction pathways. Also, the P-R amplitude difference highlights individual variations in atrial depolarization and ventricular depolarization dynamics [44]. These insights help address privacy concerns by identifying critical ECG attributes that need protection. Understanding how specific ECG features contribute to re-identification allows for the development of more secure and transparent biometric systems, safeguarding individual privacy while utilizing ECG data for clinical and research purposes.

4.3 Conclusion

This study comprehensively analyzed the re-identification risks associated with ECG data using traditional statistical features and transparent machine learning models. By validating our approach across five diverse datasets, we demonstrated that ECG signals contain sufficient biometric information to significantly compromise privacy, achieving high accuracy in re-identifying individuals. Through SHAP analysis, we identified the most impactful features, providing critical insights for clinical experts and guiding the development of effective anonymization techniques. These findings highlight the urgent need for robust privacy-preserving mechanisms to safeguard patient biosignal data in real-world health applications.

Chapter 5

TransECG: Interpreting Biometric Identity Leakage in ECG Signals via Vision Transformers

This chapter is a reprint of material published in the proceedings of the 22nd IEEE-EMBS International Conference on Body Sensor Networks (BSN 2026).

5.1 Introduction

Electrocardiogram (ECG) signals support cardiovascular diagnosis, continuous monitoring, wellness and athletic sensing, and biometric authentication [129]. Their morphology, including the P–R interval, QRS complex, ST segment, and T wave, exhibits stable variation across individuals and demographic groups [121]. While this makes ECG valuable for personalized health systems, it also means that ECG streams can reveal sensitive identity information. As wearable and clinical ECG data are shared across digital-health pipelines, re-identification

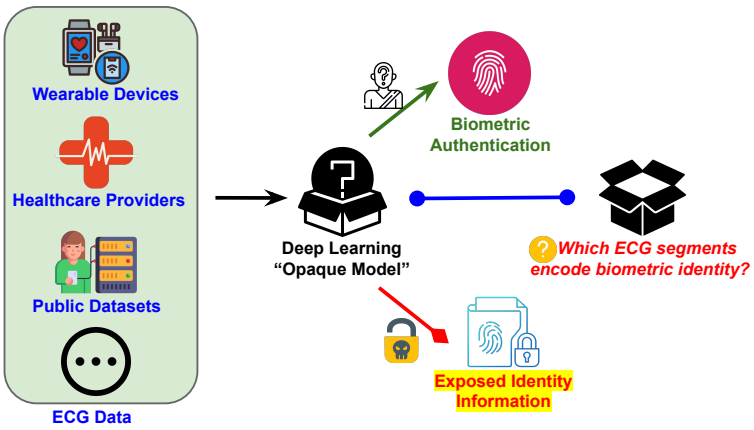


Figure 5.1: ECG data support digital-health applications but can also leak biometric identity. TransECG probes how transformer models encode identity-related ECG regions.

and linkage attacks remain barriers to open data sharing [53, 135, 71, 138].

Deep learning has improved ECG inference and identification [137], and transformer architectures are increasingly used in ECG foundation models because they capture long-range temporal dependencies [69, 81]. Yet these models remain opaque: they may expose identity, but the ECG regions driving leakage are unclear. Understanding where identity is encoded is therefore essential for trustworthy sensing.

We study how transformer-based models capture biometric identity from long-duration ECG, focusing on interpretability rather than proposing a new foundation model. Vision Transformers represent inputs as patch sequences and learn global representations through self-attention [36]; this fits ECG identity because it appears through repeated morphology across cardiac cycles rather than isolated samples. Figure 5.1 illustrates the motivation. Our contributions are: (i) a systematic study of ECG identity leakage with a transformer architecture; (ii) TransECG, an attention-based pipeline mapping model focus to physiological ECG components; and (iii) evaluation on four real-world datasets showing strong re-identification performance and task-specific identity cues.

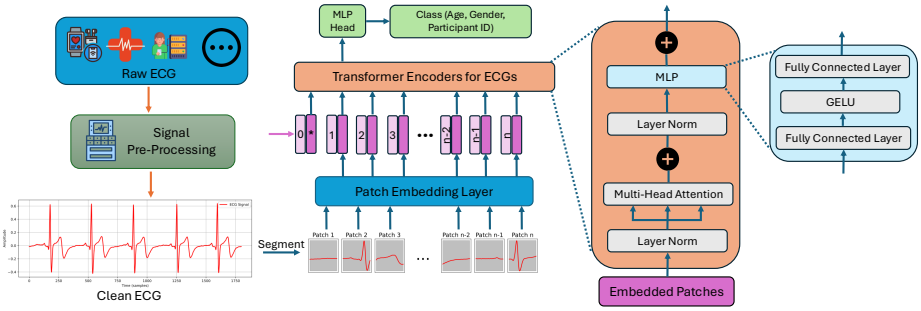


Figure 5.2: Overview of TransECG. Filtered ECG sequences are divided into temporal patches, encoded by a ViT-style transformer, and classified with a task-specific head. Final-layer class-token attention is mapped back to ECG morphology for interpretation.

5.2 Method

TransECG analyzes how transformer models encode biometric identity in long-duration ECG. As shown in Figure 5.2, it includes preprocessing, patch embedding, transformer encoders, a task-specific classifier, and attention-based interpretation. The goal is to probe a ViT-style architecture aligned with modern ECG representation learning, not to introduce a new foundation model.

5.2.1 Patch-based Modeling of Long-duration ECG

Raw ECG signals are preprocessed for quality and cross-dataset consistency. We apply a Butterworth bandpass filter to remove baseline wander and high-frequency noise, reduce motion artifacts, resample to a uniform frequency, and normalize amplitudes while preserving physiologically meaningful waveforms.

Conventional time-series transformers that tokenize individual samples can be inefficient on long ECG recordings because of excessive token length and diffuse attention [23]. Biometric identity is also not encoded at isolated time points, but emerges from recurring morphology across cardiac cycles. TransECG therefore divides each preprocessed sequence into fixed-

length, non-overlapping temporal patches. These computational units can include portions of the P wave, QRS complex, or T wave.

Given patch $\mathbf{x}_i$, TransECG flattens and projects it into a D-dimensional embedding:

$$\mathbf{z}_i = \text{Flatten}(\mathbf{x}_i)\mathbf{E}, \quad i = 1, \ldots, N,$$

where $\mathbf{E}$ is a learnable projection matrix. A learnable class token and positional embeddings are added to preserve temporal order:

$$\mathbf{Z}_0 = [\mathbf{z}_0; \mathbf{z}_1; \ldots; \mathbf{z}_N] + \mathbf{E}_{\text{pos}}.$$

The sequence is processed by L transformer encoder layers with multi-head self-attention (MHSA) and feedforward networks (FFN):

$$\mathbf{Z}'_\ell = \text{LN}(\mathbf{Z}_{\ell-1} + \text{MHSA}(\mathbf{Z}_{\ell-1})),$$
$$\mathbf{Z}_\ell = \text{LN}(\mathbf{Z}'_\ell + \text{FFN}(\mathbf{Z}'_\ell)).$$

For each attention head h, attention is computed from query, key, and value matrices as

$$\mathbf{A}_h = \text{softmax}\left(\frac{\mathbf{Q}_h \mathbf{K}_h^\top}{\sqrt{D_h}}\right)\mathbf{V}_h.$$

The output class token is passed to a task-specific MLP head for gender, age-group, or participant-ID classification. This design preserves a direct mapping between token positions and ECG time regions, enabling model behavior to be analyzed over long temporal contexts.

5.2.2 Attention-based Identity Pattern Analysis

We analyze how the transformer distributes importance across ECG signals when performing re-identification tasks. Rather than treating attention as a definitive causal explanation, we use it as a consistent model-intrinsic signal for probing how identity-related information is distributed across temporal regions. In the final transformer layer, the class token aggregates information from all patch tokens. For attention head h, the class-to-patch weights are

$$\mathbf{a}_h = \mathbf{A}_h[0, 1 : N], \qquad \text{Attention}_i = \frac{1}{H} \sum_{h=1}^{H} a_{h,i}.$$

Patch scores are projected back to the ECG timeline and aligned with physiological components, including the P wave, P–R interval, QRS complex, ST segment, Q–T interval, and T wave. Aggregating scores across samples and cycles reveals whether re-identification relies on high-amplitude QRS components, conduction intervals such as P–R, or broader repolarization patterns, without claiming attention alone establishes causality.

5.3 Experimental Results

5.3.1 Dataset and Experimental Setup

We evaluated TransECG on gender classification, five-class age-group classification, and participant ID re-identification. Age groups are 0–18, 19–35, 36–50, 51–65, and 66+ years. Four ECG databases were merged: MIT-BIH Arrhythmia [85], MIT-BIH Long-Term ECG [56], CHF Congestive Heart Failure [11], and the Brno University of Technology ECG Signal Database [11]. The final dataset contains 87 participants with diverse ages, genders, and cardiac conditions.

For preprocessing, a Butterworth bandpass filter (0.5–40 Hz) was applied to remove noise and baseline drift, followed by Pan–Tompkins R-peak detection. Participants with fewer than 2,000 samples were excluded to ensure consistent sequence lengths. Each ECG signal was resampled to 250 Hz, segmented into 2,000 time steps, and normalized using min–max scaling at the sequence level. The dataset was split into training (70%), validation (15%), and test (15%) sets with no participant overlap, supporting evaluation on unseen subjects. TransECG uses six transformer layers, six attention heads, hidden dimension 256, patch size 20, MLP dimension 128, AdamW with learning rate 1×10^{-4}, stochastic-depth survival probability 0.8, and early stopping over 45 epochs.

5.3.2 Baselines

We compare TransECG with ECG Unveiled [134], CNN-based [52], and LSTM-based [72] models. ECG Unveiled uses interpretable machine learning methods, including logistic regression, decision trees, and random forests, with manually engineered ECG features for re-identification risk analysis. It further applies SHAP to identify important PQRST features for gender, age, and participant identification. CNN and LSTM baselines represent commonly used deep-learning alternatives, with CNNs capturing local morphology and LSTMs modeling sequential dynamics.

5.3.3 Accuracy Assessment

Table 5.1 preserves all performance metrics. TransECG consistently achieves the best accuracy, precision, and F1-score: 0.899/0.900/0.899 for gender, 0.899/0.901/0.899 for age, and 0.886/0.889/0.882 for ID re-identification. These results outperform ECG Unveiled, CNN, and LSTM baselines, indicating that transformer attention captures identity-related ECG cues beyond manually engineered PQRST features while retaining interpretable localization.

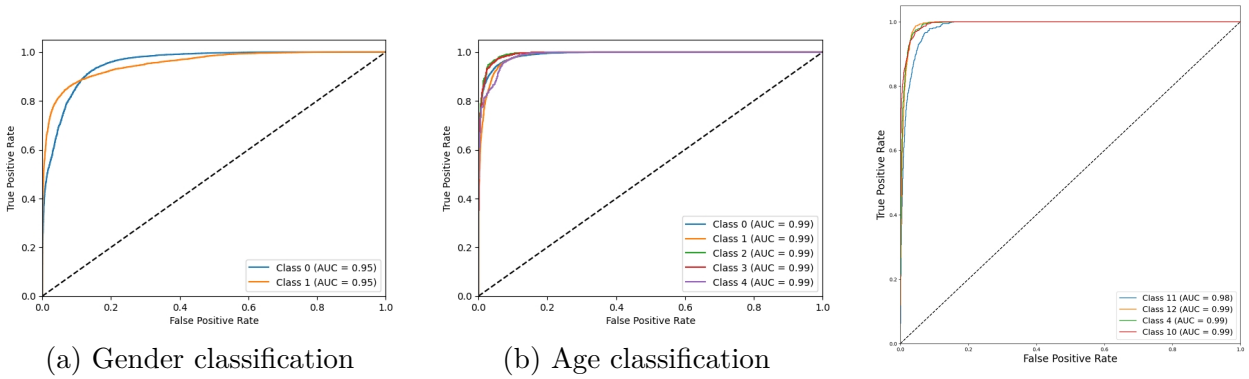
Table 5.1: Performance comparison between TransECG and baselines.

Task	**Model**	**Acc.**	**Prec.**	**F1**	**Explainability**	**Cost**	**Preprocess**
Gender	**TransECG**	**0.899**	**0.900**	**0.899**	High	High	High
	ECG Unveiled [134]	0.755	0.766	0.760	Medium	Low	Moderate
	CNN [52]	0.857	0.860	0.855	Low	Medium	Moderate
	LSTM [72]	0.863	0.870	0.865	Low	Med.-High	Moderate
Age	**TransECG**	**0.899**	**0.901**	**0.899**	High	High	High
	ECG Unveiled	0.671	0.623	0.633	Medium	Low	Moderate
	CNN	0.834	0.840	0.832	Low	Medium	Moderate
	LSTM	0.849	0.855	0.847	Low	Med.-High	Moderate
ID	**TransECG**	**0.886**	**0.889**	**0.882**	High	High	High
	ECG Unveiled	0.819	0.817	0.810	Medium	Low	Moderate
	CNN	0.842	0.845	0.840	Low	Medium	Moderate
	LSTM	0.855	0.858	0.852	Low	Med.-High	Moderate

ROC analysis further evaluates discriminative ability (Figure 5.3). AUC values are 0.95 for gender, 0.99 for age, and 0.99 for ID re-identification, with the top four ID classes displayed due to the number of participants. These scores reinforce that healthcare and wearable ECG can support reliable demographic and individual re-identification.

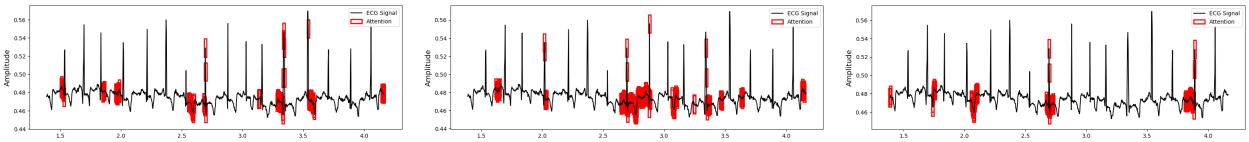
5.3.4 Attention-based Explainability

Figure 5.4 visualizes representative attention maps. For gender classification, attention concentrates on the R peak and QRS complex. The R wave is clinically meaningful because gender-related amplitude differences have been associated with heart size and ventricular mass [121]. Attention also appears on the S wave, T wave, and S–T interval, which reflect repolarization and sex-related QT differences [21], as well as P–Q/P–R conduction regions [142].



(a) Gender classification (b) Age classification (c) ID re-identification

Figure 5.3: ROC curves for (a) gender classification, (b) age classification, and (c) participant ID re-identification.



(a) Gender (b) Age (c) Participant ID

Figure 5.4: Representative attention maps for (a) gender classification, (b) age classification, and (c) participant re-identification, showing task-specific ECG patterns emphasized by the model.

For age classification, attention highlights the P–R interval, Q–T interval, and R/T-wave amplitudes. P–R emphasis is consistent with age-related slowing of atrioventricular conduction [123], while Q–T and repolarization attention reflects age-varying ventricular recovery [68, 107]. R and T wave attention also aligns with structural and electrical remodeling over aging [120, 114].

For participant ID re-identification, attention emphasizes the QRS complex, especially the R wave and Q–R/R–S segments, whose amplitude and morphology provide biometric cues linked to cardiac anatomy and conduction. The model also attends to S–T, T wave, and T–Q repolarization regions [84, 63], as well as the P wave and P–R interval [8], indicating that identity is distributed across multiple stages of the cardiac electrical cycle.

Figure 5.5 aggregates final-block attention over physiological intervals. Gender classification is driven mainly by R-wave/QRS depolarization, age classification by P–R and Q–T con-

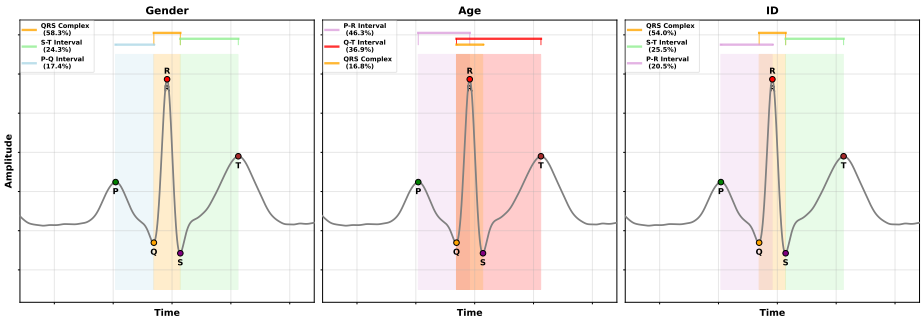


Figure 5.5: ECG regions consistently emphasized across gender, age, and participant ID re-identification tasks.

duction/recovery intervals, and ID re-identification by QRS morphology plus repolarization and conduction intervals. Compared with handcrafted baselines, TransECG dynamically identifies morphology-aware and long-range cues, including S–T and recovery regions that may be underused in manual pipelines.

5.4 Discussion

TransECG provides strong predictive performance and task-specific interpretability for ECG identity leakage. The attention patterns bridge accuracy-driven ECG modeling and transparent privacy analysis, which is important for personalized health ecosystems where data may be shared among patients, clinicians, vendors, and research platforms. The findings also show that privacy-preserving ECG sharing should go beyond metadata removal: even without explicit identifiers, morphology-level cues in QRS, conduction, and repolarization regions can support re-identification. This is especially relevant to long-duration intelligent sensing, where richer temporal context improves utility but may amplify identity leakage.

Limitations remain. Although we use four real-world datasets, the cohort size is moderate, and broader population diversity is needed. Attention-based interpretability is informative but not causal. Future work will use larger multi-institutional and wearable datasets, com-

bine attention with attribution and causal methods, and develop privacy-preserving transformations that reduce re-identification while preserving clinical utility.

5.5 Conclusion

We presented TransECG, a ViT-based framework for analyzing ECG re-identification risks with attention-based evidence. Across gender, age, and participant ID tasks, TransECG achieves strong performance while localizing identity cues to physiological ECG components. The results show that long-duration ECG can encode demographic and individual identity through stable morphology across cardiac cycles. By combining performance with structured interpretability, TransECG supports privacy-conscious ECG data sharing and trustworthy digital-health sensing.

Chapter 6

Linkage Attacks Expose Identity Risks in Public ECG Data Sharing

This chapter is based on material published in the proceedings of the 2025 47th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC).

6.1 Introduction

Electrocardiograms (ECG) capture the heart's electrical activity, serving as a key diagnostic tool for conditions like heart failure and arrhythmias [108, 7]. Beyond clinical use, ECG signals exhibit unique morphological and temporal patterns influenced by heart anatomy, conduction pathways, and autonomic regulation, making them suitable for biometric identification [52]. Unlike fingerprints, ECG signals are dynamic and vary naturally due to physiological changes, enabling intrinsic condition detection, which reduces susceptibility to spoofing and enhances security in authentication systems [91]. As telehealth platforms and

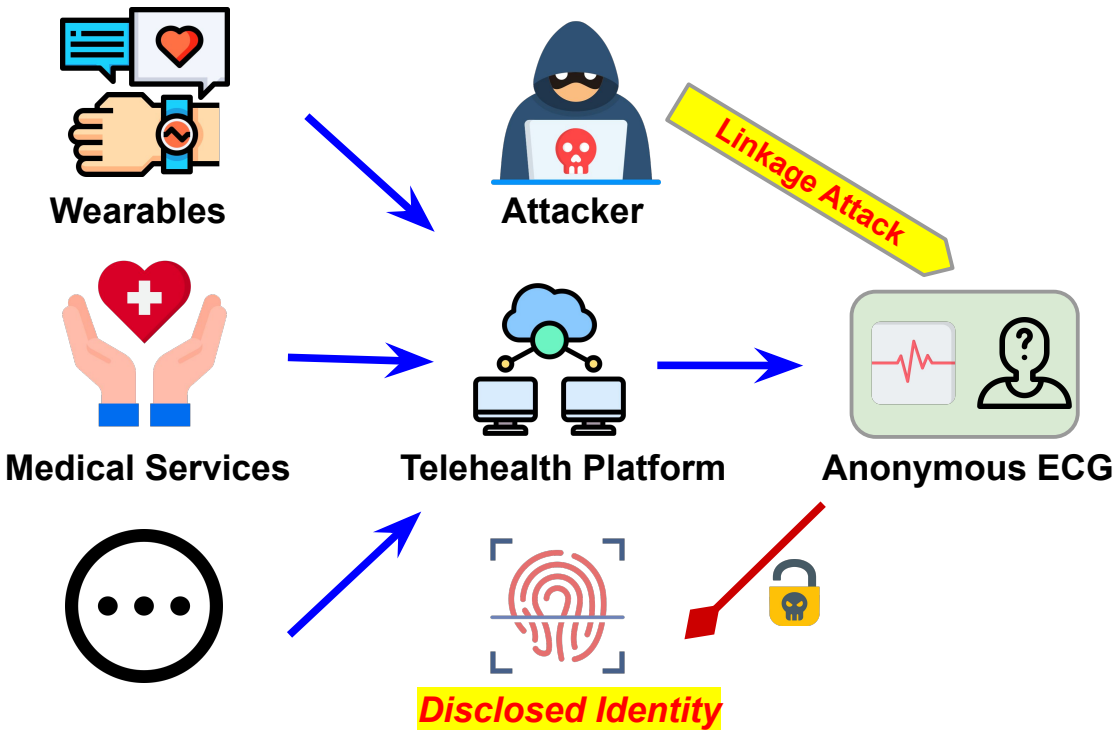


Figure 6.1: Illustration of a linkage attack on public anonymous ECG data. A telehealth platform integrates data from multiple sources, including wearable devices and medical institutions. An attacker with partial knowledge cross-references public and private datasets to re-identify individuals, compromising privacy and trust in medical data sharing.

wearable devices integrate ECG data for remote monitoring [3, 4], the increasing availability of publicly shared datasets raises significant privacy concerns [134].

Due to their biometric nature, ECG signals are vulnerable to linkage attacks, where adversaries exploit overlaps between public and private datasets to re-identify individuals. Unlike demographic data, ECG retains distinctive patterns over time, making de-identification challenging. An attacker with partial knowledge can aggregate ECG samples from wearables, telehealth platforms, or leaked records and cross-reference them with public datasets using machine learning [13]. Leveraging membership inference [113] and record linkage techniques, attackers can link anonymized ECGs to identities, compromising patient privacy. This risk grows as ECG datasets proliferate across research, clinical, and commercial domains without adequate safeguards.

Figure 6.1 demonstrates this problem, showing how telehealth platforms consolidate ECG data from diverse sources, creating a rich repository that attackers can exploit [140, 147]. By correlating leaked or externally obtained ECG signals with public datasets, adversaries can disclose identities, leading to privacy breaches. Beyond compromising confidentiality, linkage attacks undermine trust in medical data sharing, discourage research participation, and threaten the security of digital health infrastructures [141]. As healthcare increasingly relies on open data initiatives, mitigating these risks is essential to preserving patient privacy and trust [134, 136].

Despite these risks, existing ECG privacy studies often make overly simplified assumptions about adversarial capabilities [75], assuming full data access or controlled re-identification settings. In reality, attackers operate with partial and noisy knowledge, handling heterogeneous datasets from varying devices, medical conditions, and recording environments. Moreover, prior studies struggle to distinguish known from unknown individuals [52, 94], inflating success rates in controlled settings while overlooking real-world challenges like within-person variation due to aging, disease progression, or sensor differences [139]. These gaps hinder a true understanding of ECG privacy risks and the development of effective protective measures.

To address ECG privacy risks, we propose the first framework to assess re-identification under realistic linkage attacks. Unlike prior studies assuming full adversarial knowledge, we model a practical threat where attackers with partial knowledge cross-reference auxiliary ECG data with public repositories to link identities, simulating real-world risks from wearables, telehealth, and leaked medical records. We implement a Vision Transformer (ViT)-based model [36] to assess how well attackers can match ECG signals while correctly distinguishing unknown individuals. Our study provides an accurate evaluation of linkage attack risks in biosignal data sharing, demonstrating the urgent need for stronger privacy-preserving measures in healthcare.

Our contributions are as follows:

- We define and formalize ECG linkage attacks, illustrating how adversaries exploit overlaps between public and private datasets for re-identification.
- We introduce the first realistic attack model that quantifies re-identification risks under partial adversarial knowledge, aligning with real-world telehealth and data-sharing constraints.
- We provide empirical evidence of ECG privacy vulnerabilities across diverse datasets, establishing benchmarks to guide privacy-preserving research.

6.2 Methodology

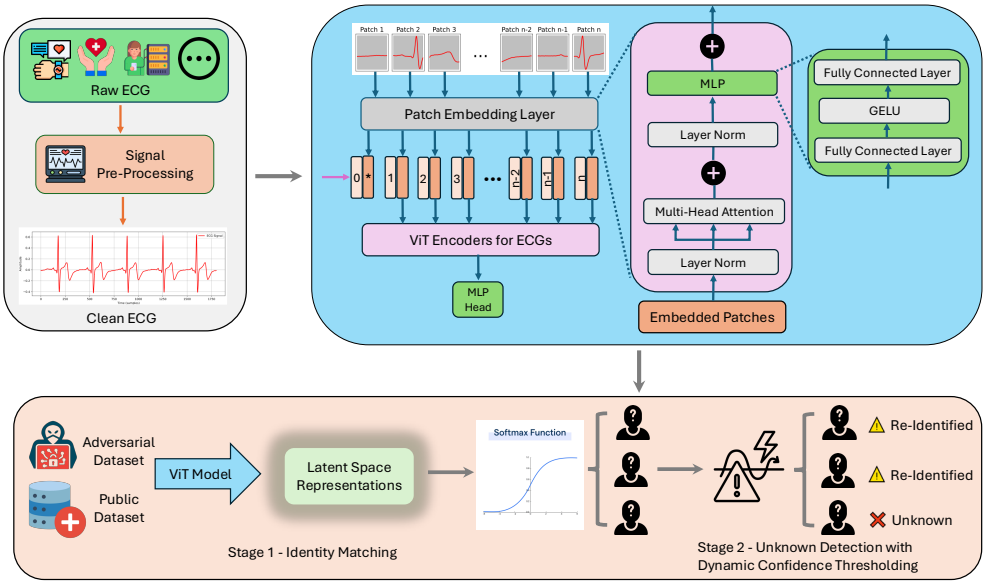


Figure 6.2: Overview of the proposed linkage attack method.

The biometric nature of ECG signals makes publicly shared datasets vulnerable to linkage attacks, where adversaries exploit overlaps between public and private data to re-identify individuals. Figure 6.2 illustrates our ViT-based framework. First, raw ECG signals undergo

pre-processing, segmentation, and embedding. A ViT encoder extracts identity-specific features, followed by an Multi-Layer Perceptron (MLP) classifier. Stage 1 (Identity Matching) leverages an adversarial dataset—collected from public or leaked sources—to train a ViT model that maps ECG signals to latent space representations, assigning identity probabilities via softmax. Stage 2 (Unknown Detection) applies a dynamic confidence threshold to reject uncertain matches, reducing false identifications. This framework quantifies re-identification risks, showing that adversaries can link ECG data even with partial knowledge, underscoring the need for stronger privacy safeguards.

6.2.1 Problem Definition

Let $\mathcal{D}_{\text{pub}} = \{(x_i, y_i)\}_{i=1}^{N}$ be a public ECG dataset, where $x_i \in \mathbb{R}^T$ is an ECG signal of length T, and $y_i \in \mathcal{Y}$ is the corresponding identity label. Similarly, let $\mathcal{D}_{\text{priv}} = \{x_j\}_{j=1}^{M}$ be a private dataset containing ECG signals from both *known participants*, whose identities exist in $\mathcal{D}_{\text{pub}}$, and *unknown participants*, who are absent from $\mathcal{D}_{\text{pub}}$. Given $x_j \in \mathcal{D}_{\text{priv}}$, an attacker seeks to assign a label $\hat{y}_j \in \mathcal{Y} \cup \{\text{unknown}\}$, identifying whether the signal belongs to a known or unseen individual.

We investigate linkage attack risks, where an adversary attempts to re-identify individuals by cross-referencing ECG samples in $\mathcal{D}_{\text{priv}}$ with identities in $\mathcal{D}_{\text{pub}}$. The attacker uses an adversarial dataset—ECG signals collected from public sources, leaked medical data, or unauthorized means, containing identifiable records. Unlike standard classification, the attacker operates under partial knowledge, meaning they do not know whether a given identity exists in $\mathcal{D}_{\text{pub}}$. This introduces practical challenges, such as dataset mismatches, ECG variability across devices and conditions, and the need to differentiate between known and unknown individuals. Our study formalizes this threat model and quantifies the re-identification risks posed by such attacks in real-world data-sharing scenarios.

6.2.2 Framework for Linkage Attack Evaluation

The proposed framework evaluates re-identification risks by simulating an attacker attempting to link ECG samples from $\mathcal{D}_{\text{priv}}$ to individuals in a black-box public dataset $\mathcal{D}_{\text{pub}}$. Unlike traditional attacks that assume full adversarial knowledge [134, 52], our framework simulates real-world constraints, where the attacker has limited visibility and no prior knowledge of the overlap between $\mathcal{D}_{\text{priv}}$ and $\mathcal{D}_{\text{pub}}$. The attacker employs a machine learning-based attack using a ViT, training on an adversarial dataset (ECG data collected from public records, leaks, or unauthorized sources) without explicit knowledge of which identities exist in the target dataset.

6.2.3 ViT-Based Representation Learning

To effectively encode ECG signals, the ViT-based model [36] applies self-attention mechanisms to capture both local and global dependencies in the signal. Unlike convolutional architectures [75], which primarily rely on spatial locality, the ViT capture long-range dependencies, making them well-suited for biosignals.

Patch Embedding and Positional Encoding Given an ECG sequence $x_i \in \mathbb{R}^T$ from $\mathcal{D}_{\text{priv}}$, we divide it into fixed-length patches of size P, resulting in $N = T/P$ patches. Each patch is linearly projected into a d-dimensional space:

$$z_0 = W_p x_i + b_p$$

where $W_p \in \mathbb{R}^{d \times P}$ is a learnable projection matrix. A learnable classification token z_{cls} is prepended to the sequence, and positional encodings E are added to retain temporal

structure:

$$Z = [z_{\text{cls}}, z_0, z_1, ..., z_N] + E$$

This embedding is passed through multiple transformer layers for feature extraction.

Multi-Head Self-Attention Mechanism Each transformer layer consists of a multi-head self-attention module that computes relationships between input patches across multiple attention heads. The attention scores for each head are computed as:

$$A_h = \text{softmax}\left(\frac{Q_h K_h^T}{\sqrt{d_k}}\right)$$

where Q_h, K_h, V_h are the query, key, and value matrices for head h, and d_k is the scaling factor. The outputs from all heads are concatenated and linearly transformed:

$$Z' = \text{Concat}(A_1 V_1, A_2 V_2, ..., A_H V_H) W_O$$

where H is the number of attention heads, and W_O is a learnable projection matrix. This is followed by layer normalization and a feedforward network. The final feature representation is extracted from z_{cls} and passed to a classifier, encoding the ECG signal into a latent feature space where identity-specific characteristics are captured.

6.2.4 Two-Stage Classification for Known and Unknown Participants

With latent space representations from the ViT, the attacker seeks to match an ECG sample to an identity in $\mathcal{D}_{\text{pub}}$. However, without prior knowledge of the target dataset's composition, direct classification is uncertain, risking false matches for unseen individuals. Notably, this attack remains practical—the attacker trains a model solely on their adversarial dataset,

without external knowledge or privileged access. Rather than relying on handcrafted rules, the model autonomously learns patterns, demonstrating that re-identification is feasible even with limited adversarial knowledge.

To address this challenge, the attack follows a two-stage classification process:

Stage 1: Identity Matching The ViT model processes the input ECG sample and outputs a probability distribution over the set of known identities:

$$P(y|x_j;\theta) = \text{softmax}(Wz + b),$$

where W and b are trainable parameters. The identity with the highest probability is selected as the predicted label:

$$\hat{y}_j = \arg\max P(y|x_j;\theta).$$

This step assumes that if the sample belongs to an individual already present in $\mathcal{D}_{\text{pub}}$, the model will confidently assign the correct identity. However, if the sample comes from an unknown participant, the model may still attempt to match it to the closest known identity, leading to misclassification.

Stage 2: Unknown Detection with Dynamic Confidence Threshold Since the attacker lacks full knowledge of $\mathcal{D}_{\text{pub}}$, it is critical to determine when a test sample does not match any known identity. To mitigate false positive matches, we introduce a dynamic confidence threshold Φ, which adapts based on the confidence scores of known samples:

$$\Phi = \text{percentile}(\{\tau(x) \mid x \in \mathcal{D}_{\text{pub}}\}, p),$$

where $\tau(x)$ is the confidence score of the highest prediction, and p is a tunable percentile parameter. This means that the threshold is determined by the confidence levels observed

in the public dataset, making it adaptable to different distributions of ECG signals.

If the model's confidence in its highest prediction for x_j is below Φ, the sample is classified as *unknown*, rather than being forcefully assigned to a known identity. This approach allows the attack to selectively reject uncertain predictions, reducing the risk of misidentifying individuals who are not in $\mathcal{D}_{\text{pub}}$. By dynamically adjusting Φ, the model balances sensitivity and specificity, improving the robustness of the attack. Despite these constraints, this attack remains effective because the attacker leverages only their own independently trained model, making no assumptions about external datasets, manually crafted decision rules, or privileged knowledge about $\mathcal{D}_{\text{pub}}$.

6.2.5 Design Considerations and Practical Constraints

Our framework simulates real-world adversarial conditions, ensuring the attack reflects practical constraints rather than an idealized setting. Since ECG signals encode both biometric and physiological variations, the ViT model effectively extracts distinguishing features while handling variability. A dynamic confidence threshold mitigates false matches by preventing identity assignments under high uncertainty. This study provides a realistic assessment of privacy risks in publicly shared ECG data, highlighting the limitations of traditional anonymization techniques like noise addition, which often degrade diagnostic value. Alternative privacy-preserving strategies, such as differential privacy [141] and federated learning [31], are needed to mitigate re-identification risks while preserving ECG utility.

6.3 Experiments and Results

We conduct a comprehensive evaluation of our framework for ECG linkage attack analysis, focusing on re-identification risks, robustness under adversarial settings, and benchmarking

against state-of-the-art methods. Additionally, we analyze the impact of different data splits and perform an ablation study to assess key framework components.

6.3.1 Datasets

To ensure a robust evaluation, we utilize publicly available ECG datasets covering diverse demographics, recording conditions, and health states, summarized in Table 6.1. Each dataset undergoes preprocessing, including resampling to a consistent sampling rate, segmentation into fixed-length windows (e.g., 10s intervals), and z-score normalization.

Table 6.1: Summary of ECG Datasets Used in Experiments

Dataset	**Subjects**	**Age range**	**Sex (M/F)**	**Rate (Hz)**	**Health condition**
MIT-BIH Arrhythmia [85]	47	23–89	25/22	360	Arrhythmias
MIT-BIH Malignant Ventricular [57]	22	22–78	–	250	Ventricular tachycardia
BIDMC CHF [11]	15	22–71	11/4	250	Congestive heart failure
Brno University of Technology ECG Quality [89]	15	21–83	9/6	1000	General population
MIT-BIH Long-Term [56]	7	46–88	6/1	128	Long-term monitoring
Combined	**106**	**21–89**	**Varies**	**Varies**	**Multiple**

6.3.2 Experimental Setup

Adversarial Scenarios To evaluate the robustness of our framework, we simulate realistic attack conditions under different levels of adversarial knowledge. The attacker's goal is to re-identify individuals in the public dataset $\mathcal{D}_{\text{pub}}$ using an adversarial dataset $\mathcal{D}_{\text{priv}}$. However, the attacker does not know in advance whether an individual in $\mathcal{D}_{\text{priv}}$ exists in $\mathcal{D}_{\text{pub}}$. We consider the following scenarios:

- **Partial Knowledge (Realistic Scenario):** The attacker trains on $\mathcal{D}_{\text{priv}}$, which includes ECG samples from individuals who may or may not exist in $\mathcal{D}_{\text{pub}}$. Without prior knowledge of dataset overlap, the attacker relies on independently collected data to attempt re-identification, reflecting real-world constraints.

- **Full Knowledge (Upper Bound):** The attacker has full access to $\mathcal{D}_{\text{pub}}$ during training, ensuring all target individuals are present. This reduces the task to standard classification rather than true re-identification, representing an unrealistic but best-case scenario for attack performance.

- **Noisy Data (Robustness Test):** To evaluate the framework's resilience, we introduce synthetic noise into $\mathcal{D}_{\text{pub}}$, simulating real-world distortions such as sensor errors or variations in ECG recordings. This tests the model's ability to generalize under imperfect data conditions.

Data Splits To systematically assess re-identification risks, we design training-validation-testing splits that reflect practical adversarial constraints. The attacker's dataset is constructed by sampling from $\mathcal{D}_{\text{pub}}$, but the attacker does not know in advance which individuals are present in the test set. Specifically, we divide $\mathcal{D}_{\text{pub}}$ into three subsets:

- **Known Individuals:** Participants whose ECG data appears in both $\mathcal{D}_{\text{pub}}$ and the attacker's training set. The attacker learns to classify these individuals but does not have access to their full ECG sequences.

- **Unknown Individuals:** Participants in $\mathcal{D}_{\text{pub}}$ whose data is completely absent from the attacker's training set. The attacker must recognize that these individuals are not part of the known set and classify them as *unknown* rather than incorrectly linking them to an existing identity.

- **Test Set:** A mix of known and unknown individuals used to evaluate the model's ability to correctly match identities while minimizing false positives and false negatives.

We employ multiple data splits (e.g., 50/25/25, 60/20/20, and 70/15/15 for training, validation, and testing) to assess how different levels of data exposure influence attack success. These splits simulate a practical attack scenario where the adversary is trained only on a subset of known individuals and must distinguish between known and unknown participants in the test phase. Since the attacker does not have full visibility into the dataset, they must generalize from limited observations, making this a realistic evaluation of re-identification risks in public ECG data sharing.

Evaluation Metrics To comprehensively assess model performance, we employ the following evaluation metrics:

- **Sample-Level Metrics:** Accuracy, precision, recall, and F1-score measure how well individual ECG samples are correctly identified. Higher precision indicates fewer false positives, while higher recall reflects fewer missed identifications. These metrics capture the reliability of ECG-based identity matching.

- **Participant-Level Metrics:** Re-identification rate measures how often the model correctly links individuals from the adversarial dataset to the target dataset. Protection rate quantifies how well the system prevents unknown individuals from being misclassified as known identities. These metrics assess identity leakage risks in practical settings.

- **Unknown Detection Metrics:** True Negative Rate (TNR) measures the ability to correctly reject unknown individuals, while False Positive Rate (FPR) captures how often they are mistakenly linked to known identities. Confidence thresholds regulate the decision boundary, helping the model avoid overconfident misclassifications in real-

world deployments.

- **Threshold-Based Metrics:** Since confidence thresholds critically impact classification performance, we evaluate Equal Error Rate (EER) [90], defined as the threshold where the False Acceptance Rate (FAR) equals the False Rejection Rate (FRR):

$$\text{EER} = \arg\min_{\theta} |\text{FAR}(\theta) - \text{FRR}(\theta)|$$

  To compute EER, the threshold is varied to generate a Receiver Operating Characteristic (ROC) curve, with EER located where FAR and FRR intersect. A lower EER indicates better privacy preservation by reducing both false matches and missed identifications, balancing identity protection with classification accuracy.

Data Processing To ensure consistency across datasets, ECG signals are preprocessed through standardization and segmentation. The raw signals are first resampled to a target frequency of 250 Hz to unify different sampling rates. Following this, each signal is normalized using min-max scaling and divided into fixed-length segments of 2000 samples per window to maintain uniform input sizes for the model. A label encoding process is applied to standardize participant identities across datasets. Additionally, synthetic data augmentation is incorporated to introduce variations through transformations such as additive noise, signal scaling, polarity flipping, and temporal shifting [117]. This processing pipeline maintains the integrity of the ECG signals while making the model robust to real-world variations in biosignals.

Implementation Details The attack model employs a ViT to extract identity-specific representations from ECG signals, processing sequences with a patch size of 20, an embedding dimension of 256, and six transformer layers with eight self-attention heads. Each layer includes an MLP of dimension 128, with stochastic depth regularization (0.8 survival

probability) for stability. The model is trained using AdamW with a learning rate of 0.0001, weight decay of 10^{-4}, and a cosine schedule, with early stopping based on validation F1-score. Training runs for 300 epochs with a batch size of 64 on an NVIDIA RTX 4090 GPU, using data augmentation techniques such as jittering, cropping, and scaling. The attack follows a two-stage classification, where a discriminator first determines if an ECG sample belongs to a known identity before ViT refines the classification. The discriminator is a fully connected network with a hidden dimension of 128, ReLU activation, batch normalization, and dropout (0.3). Overconfident misclassifications are mitigated using a dynamic confidence threshold.

6.3.3 Results and Analysis

Baseline Comparison Table 6.2 compares our framework against state-of-the-art models. Our approach consistently outperforms traditional machine learning methods across all evaluation metrics. Accuracy represents the proportion of correctly classified samples. Precision measures the fraction of correctly predicted positive instances among all predicted positives, indicating the model's reliability in classification. Recall quantifies the proportion of actual positives correctly identified, assessing the model's sensitivity. F1-score is the harmonic mean of precision and recall, providing a balanced measure of classification performance. EER represents the threshold at which the FAR equals the FRR, meaning the probability of misclassifying an unknown individual as a known participant is equal to the probability of failing to recognize a known participant. A lower EER enhances privacy by minimizing both false identifications and misclassifications. Our ViT-based approach achieves an EER of 0.127, significantly lower than traditional machine learning models such as Random Forest (0.232) and Logistic Regression (0.289). This result confirms the model's robustness in preventing unauthorized identity matching, reinforcing its effectiveness in linkage attacks.

Table 6.2: Baseline Comparison on Sample-Level Metrics.

Model	**Accuracy**	**Precision**	**Recall**	**F1-Score**	**EER**
Random Forest [134]	0.54	0.58	0.56	0.57	0.232 ± 0.006
Logistic Regression [94]	0.33	0.34	0.30	0.32	0.289 ± 0.005
XGBoost [132]	0.60	0.60	0.59	0.60	0.198 ± 0.007
CNN [52]	0.72	0.71	0.72	0.72	0.151 ± 0.004
Ours	**0.85**	**0.83**	**0.82**	**0.83**	$\mathbf{0.127 \pm 0.003}$

Impact of Data Splits Table 6.3 examines the effect of different training splits on the model's performance under the partial knowledge attack scenario. As the proportion of training data increases, the model achieves better generalization, leading to higher scores. The best performance is observed in the 70/15/15 split, where a larger training set provides the model with a more diverse representation of known individuals, improving classification capability. However, increasing the training proportion may also raise re-identification risks in real-world scenarios. Attackers often accumulate ECG data from wearables, telehealth services, or leaked medical records, gradually improving their ability to match anonymized signals to individuals. This underscores the need for limiting public data exposure and implementing privacy-preserving techniques such as differential privacy to mitigate these risks.

Table 6.3: Performance Under Different Data Splits (Partial Knowledge).

Data Split	**Accuracy**	**Precision**	**Recall**	**F1-Score**
60/20/20	0.71	0.95	0.70	0.79
50/25/25	0.72	0.96	0.71	0.80
70/15/15	0.74	0.97	0.73	0.82

Adversarial Scenario Analysis Table 6.4 compares model performance under different levels of adversarial knowledge. As expected, the full knowledge attack scenario achieves the highest performance across all metrics, demonstrating the privacy risks associated with fully exposed datasets. If an attacker gains unrestricted access to an ECG dataset, the likelihood of identity disclosure increases significantly, highlighting the importance of implementing

access control mechanisms in public health repositories.

In contrast, the partial knowledge scenario represents a more realistic setting where the attacker has limited prior information. While the attack success rate is lower than in the full knowledge scenario, the model still achieves a high F1-score of 0.82, suggesting that even limited exposure to known identities enables significant re-identification potential.

The noisy data scenario simulates real-world distortions, such as sensor errors and preprocessing variability. Despite a slight performance drop, the model maintains a notable re-identification capability (F1-score = 0.70), indicating that simple noise-based anonymization is ineffective. More robust privacy-preserving techniques, including federated learning, differential privacy, and cryptographic methods, are needed to mitigate adversarial inference.

These findings demonstrate that linkage attacks on ECG data pose a persistent privacy risk, even under limited adversarial knowledge and signal distortions. This underscores the necessity of privacy-aware data-sharing strategies, such as minimizing public exposure, enhancing anonymization, and enforcing access controls to reduce re-identification risks.

Table 6.4: Performance Under Adversarial Scenarios.

Scenario	**Accuracy**	**Precision**	**Recall**	**F1-Score**
Partial Knowledge	0.74	0.97	0.73	0.82
Full Knowledge	0.86	0.84	0.83	0.84
Noisy Data	0.72	0.70	0.71	0.70

Misclassification rates of the Attack Table 6.5 presents the misclassification rates of different models in the linkage attack. The FPR represents the proportion of unknown individuals mistakenly classified as known, indicating a privacy risk, while the FNR quantifies the proportion of known individuals misclassified as unknown, reflecting the attack's effectiveness in re-identifying participants. The overall misclassification rate aggregates both errors, providing a holistic measure of model reliability.

Traditional machine learning models, such as Random Forest and Logistic Regression, exhibit high FPR (0.28 and 0.35, respectively), frequently misidentifying unknown participants as known, significantly increasing the risk of privacy breaches. In contrast, deep learning models, including CNN and our ViT-based approach, achieve substantially lower FPR (0.15 and 0.12, respectively), indicating better resistance to erroneous re-identifications. Notably, the FNR remains relatively low for CNN (0.20) and ViT (0.18), highlighting the enhanced ability of deep learning methods to re-identify known individuals while reducing misclassification of unknown participants.

Our approach achieves the best trade-off between privacy preservation and attack efficacy, with an FPR of 0.12, FNR of 0.18, and an overall misclassification rate of 0.15. These results indicate that while our model improves re-identification accuracy, it also reduces false matches, leading to a more precise yet privacy-conscious attack. However, this also implies that as adversaries gain access to better architectures, linkage attacks on ECG data may become increasingly effective, underscoring the necessity for stronger privacy defenses.

Table 6.5: Misclassification Rates of Linkage Attack.

Model	**FPR**	**FNR**	**Misclassification Rate**
Random Forest [134]	0.28	0.32	0.30
Logistic Regression [94]	0.35	0.38	0.36
XGBoost [132]	0.22	0.26	0.24
CNN [52]	0.15	0.20	0.18
Ours	**0.12**	**0.18**	**0.15**

Confidence-Based Misclassification To further analyze misclassification behavior, Table 6.6 reports the percentage of unknown individuals misclassified as known ($\mathbf{U} \rightarrow \mathbf{K}$) and known individuals misclassified as unknown ($\mathbf{K} \rightarrow \mathbf{U}$) at varying confidence thresholds.

Lower confidence thresholds (e.g., 0.01) result in higher misclassification rates, as the model assigns identities more aggressively, increasing privacy risks. At this threshold, 40.5% of unknown individuals were incorrectly classified as known, resulting in an overall misclassi-

fication rate of 38.1%. As the threshold increases, the model becomes more conservative, reducing both false positives and false negatives. At a threshold of 0.05, misclassification rates drop to 15.6% for U → K and 12.8% for K → U, with an overall error rate of 14.2%.

Table 6.6: Confidence-Based Misclassification in Linkage Attack.

Thres.	**U → K (%)**	**K → U (%)**	**Total (%)**
0.01	40.5	35.8	38.1
0.02	35.2	30.5	32.9
0.03	28.6	24.3	26.4
0.04	22.1	18.7	20.3
0.05	**15.6**	**12.8**	**14.2**

Note: U → K: Unknown misclassified as Known. K → U: Known misclassified as Unknown. Total: Overall misclassification rate.

These results underscore the necessity of selecting an appropriate confidence threshold to balance privacy protection and identification accuracy. Lower thresholds allow for higher attack success rates but significantly increase the risk of identity leaks, while higher thresholds mitigate privacy risks at the cost of reduced re-identification capability. Our findings suggest that thresholds between 0.04 and 0.05 offer an optimal trade-off, where the model effectively distinguishes between known and unknown individuals while limiting unnecessary identity matches.

Implications for Privacy Risks Our results show that deep learning-based linkage attacks remain effective even with partial adversarial knowledge. Lower confidence thresholds increase attack success but heighten privacy risks by misclassifying unknown individuals, while higher thresholds reduce false positives but hinder re-identification of known individuals. This highlights the limitations of confidence-based filtering as a standalone privacy measure. Effective ECG data protection requires a combination of differential privacy, access control, and encrypted computation to mitigate adversarial inference.

6.4 Discussion and Conclusion

This study systematically evaluates re-identification risks in publicly shared ECG datasets, demonstrating that linkage attacks remain effective even under partial adversarial knowledge. While the full knowledge scenario—where attackers have unrestricted access to public datasets—serves as an upper bound, the partial knowledge setting is more realistic, relying on externally collected ECG samples. Despite this constraint, our results show that attackers can still achieve high re-identification accuracy by cross-referencing external datasets with anonymized ECG records. This highlights the limitations of simple anonymization techniques, such as noise injection, and underscores the urgent need for stronger privacy-preserving strategies.

To mitigate these risks, future work should explore differential privacy, federated learning, and cryptographic techniques to limit adversarial exploitation while preserving data utility. Extending this framework to other biosignals, such as EEG and PPG, would offer broader insights into biometric privacy risks. Stricter data-sharing policies, combined with privacy-aware machine learning approaches, are essential to securing biosignal datasets against linkage attacks and promoting ethical data sharing in healthcare.

Chapter 7

Membership Inference Attacks Expose Participation Privacy in ECG Foundation Encoders

This chapter is a reprint of material published in Smart Health*, vol. 41, Art. no. 100680, 2026, doi:* https://doi.org/10.1016/j.smhl.2026.100680.

7.1 Introduction

Electrocardiograms (ECGs) are among the most widely deployed physiological modalities in connected health [38]. They are routinely collected in emergency departments and inpatient wards, and increasingly measured through ambulatory and at-home monitoring, supporting applications from rapid clinical triage and arrhythmia screening to long-term risk assessment. Beyond clinical utility, modern biosignal ecosystems are becoming increasingly *shareable and deployable at scale*: public repositories and community benchmarks such as PhysioNet [56]

and PTB-XL [131] enable reproducible research, while real-world connected-health pipelines rely on cross-institution data exchange across hospitals, vendors, and cloud platforms to improve robustness in heterogeneous populations [140, 141].

Motivated by the rapid progress of foundation models in language [93, 37, 144], biosignal modeling is undergoing a similar shift from task-specific supervised training to *foundation-style* representation learning. In this paradigm, a general-purpose encoder is pretrained on large unlabeled corpora and then reused across downstream tasks, domains, and institutions. For ECG, recent progress includes contrastive learning of cardiac signals across patients and time (e.g., CLOCS) [73], universal time-series SSL (e.g., TS2Vec) [151], and masked reconstruction objectives (e.g., MAE) [60]. Importantly, this foundation trend is not unique to ECG: large-scale wearable data have enabled open foundation models for optical physiological signals such as photoplethysmography (PPG), including Pulse-PPG [102] and PaPaGei [95], alongside industry efforts that train and deploy wearable biosignal foundation models in practice [101, 127]. Together, these developments make it increasingly common to package pretrained encoders as reusable components accessed through APIs for scoring, monitoring, and downstream personalization.

However, broad reuse of biosignal foundation encoders raises a subtle but consequential privacy question: *does the model reveal whether a specific individual contributed data to pretraining?* This is a form of participation privacy risk—membership in a training corpus can itself be sensitive, even when raw waveforms and diagnostic labels are never disclosed. In connected health, membership may imply that a person visited a particular institution, participated in a study, or belongs to a cohort associated with stigmatized conditions, creating downstream harms from discrimination to unwanted health inferences. Such concerns are amplified for biosignals because physiological traces often contain stable subject-specific morphology and can support biometric recognition [75, 135, 136]. Crucially, unresolved participation leakage can also degrade trust in clinical AI and discourage institutions and

individuals from contributing or sharing data, limiting the scale and diversity of corpora that foundation-style biosignal models rely on for generalization [58, 134].

We study this risk through membership inference attacks, which aim to determine whether a target individual was included in a model's training data. Membership inference has been extensively demonstrated in computer vision and language modeling [113, 148, 19, 88], yet systematic evidence for self-supervised and foundation-style *biosignal* encoders remains limited. This gap matters because the biosignal setting differs from canonical image/text benchmarks: training data are often cohort-based with relatively few distinct subjects, each contributing many correlated windows, making subject participation potentially more separable. At the same time, connected-health deployments rarely expose full model internals; they instead expose limited *observable interfaces* such as scalar quality or anomaly scores, or derived statistics aggregated across repeated queries. Therefore, participation risk in biosignal foundation encoders cannot be assumed from prior domains and must be measured directly under deployment-aligned access constraints. Figure 7.1 illustrates this deployment pipeline and its resulting participation-leakage surface.

A key driver of participation leakage is *what the attacker can observe*. Many realistic deployments return only scalar outputs such as reconstruction errors for quality control, anomaly scores, or similarity-derived metrics, restricting adversaries to score-only inference. Moreover, practical adversaries can often query multiple windows from the same individual and aggregate evidence across time, which is particularly relevant for ECG where recordings are long and windowed. Some connected-health systems additionally expose latent representations as "feature APIs" to support retrieval, clustering, personalization, or downstream training. Such embedding-access interfaces substantially enlarge the attack surface: representation geometry itself may encode subject-specific stability induced by pretraining. In this work, we therefore extend our audit beyond scalar-only interfaces to explicitly evaluate embedding-space membership inference, modeling deployments where feature vectors are

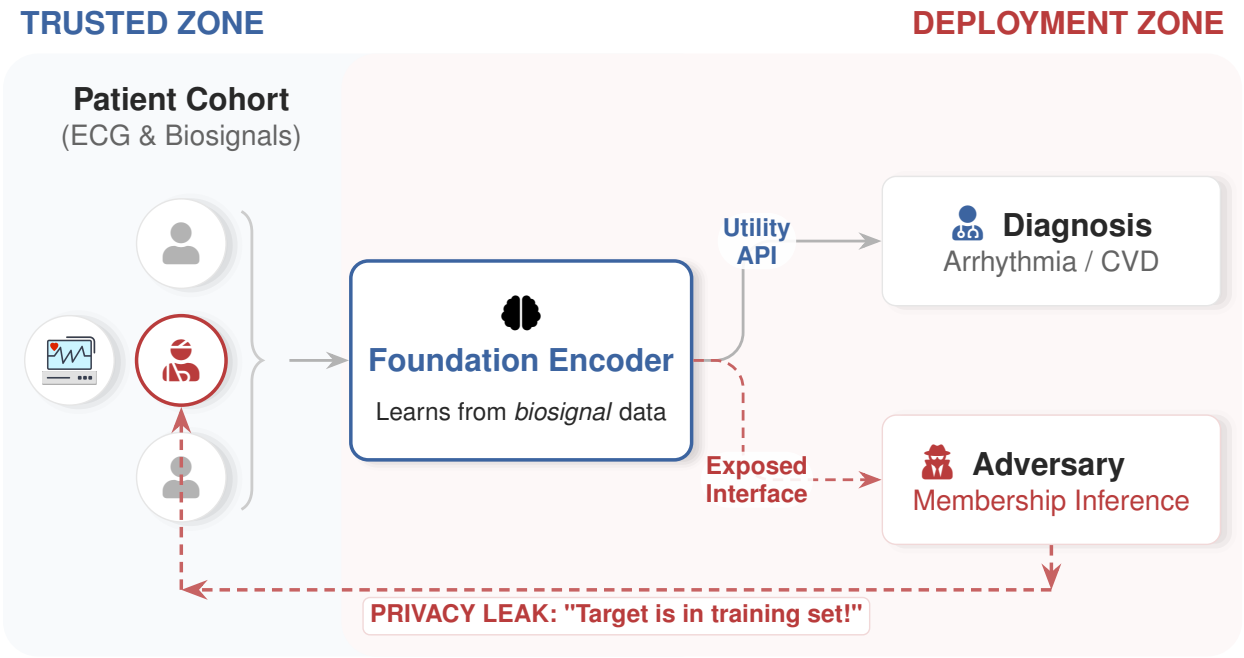


Figure 7.1: Motivation and threat surface in connected health. A foundation encoder trained on biosignal cohorts may be deployed through exposed interfaces, enabling membership inference and participation privacy leakage.

accessible to external components or partners. Our evaluation follows a subject-centric protocol with window-to-subject aggregation and calibration at fixed false-positive rates. Because real-world cohorts are often near-exhaustively used for pretraining, we audit membership using a cross-dataset non-member construction in which non-members are drawn from auxiliary ECG datasets disjoint from the training dataset. This protocol models a realistic *cross-distribution participation leakage* threat (e.g., across institutions or data sources), and may include dataset-level confounding factors (such as acquisition or resampling artifacts).

We summarize our contributions as follows.

- We present an implementation-grounded audit of membership inference against self-supervised ECG foundation encoders across multiple real-world ECG datasets and model families.

- We formalize a deployment-grounded participation privacy threat model for connected

health, organized by observable interfaces (e.g., scalar scores, aggregated statistics, and embedding access), and discuss stronger representation-access surfaces as an important future direction.

- We propose an operational, subject-centric auditing protocol calibrated at fixed FPR operating points and report standardized leakage metrics (AUC, TPR@FPR, and attacker advantage), revealing objective- and cohort-dependent participation leakage under a cross-dataset auditing protocol.

7.2 Background and Related Work

Foundation-style representation learning for ECG and biosignals. Inspired by the rapid progress of foundation models in language [93, 37, 144, 139, 137], ECG modeling is shifting from task-specific supervised learning to *foundation-style* representation learning, where general-purpose encoders are pretrained on large unlabeled corpora and reused across tasks and institutions. This shift is enabled by public resources such as PhysioNet [56] and PTB-XL [131], and accelerated by modern SSL objectives. Contrastive frameworks (e.g., SimCLR [25]) have inspired ECG-specific methods such as CLOCS [73] and general time-series SSL such as TS2Vec [151], while masked reconstruction (e.g., MAE [60]) provides an alternative route to transferable encoders. Foundation pretraining is also expanding beyond ECG to wearable biosignals, including open PPG foundation models trained across lab and field settings (Pulse-PPG) [102] and optical physiological foundation models such as PaPaGei [95], alongside large-scale industry efforts trained on wearable biosignals for deployment [101, 127]. While these models improve data efficiency and generalization, they also concentrate sensitive information into reusable latent spaces, raising privacy risks when encoders are shared widely.

Privacy and identity risks in biosignals and connected health. Privacy risks in biosignals extend beyond explicit identifiers [134, 136]. ECG waveforms capture stable subject-specific morphology and can be leveraged for biometric recognition, including deep-learning-based identification from ECG features [75, 135]. In real-world connected health systems, ECG data are often collected longitudinally and exchanged across hospitals, vendors, and cloud platforms to support monitoring, personalization, and large-scale model development [7, 140]. This ecosystem makes identity-centric threats more realistic: even after de-identification, records can be re-identified or linked across datasets through physiological signatures, repeated measurements over time, and cross-release correlations [135, 33]. At the same time, the increasing reliance on pretrained representations and reusable encoders shifts the privacy boundary from *raw signal release* to *model-level leakage*—what information can be inferred from a deployed model or its representations. This perspective aligns with broader concerns in healthcare privacy and data sharing, where participation and linkage risks can carry sensitive implications even when direct identifiers are removed [52].

Participation privacy auditing. MIAs aim to infer whether a specific record was used to train a model, providing a concrete auditing lens for *participation privacy*. Early work introduced shadow-model based MIAs against supervised classifiers [113], followed by loss-based attacks that relate membership advantage to generalization gaps [148]. Subsequent work demonstrated that MIAs can remain effective under stronger adversaries and more realistic threat models, including white-box access to model gradients and internals [88]. More recently, auditing studies have revisited membership leakage for modern large models, emphasizing that memorization and training dynamics can create measurable privacy risk even without explicit overfitting [19]. However, biosignal foundation encoders differ from canonical vision/NLP settings in important ways: datasets are often cohort-based with relatively few distinct subjects, each subject contributes many correlated windows, and deployed interfaces may expose scalar scores or latent embeddings rather than class probabilities. These

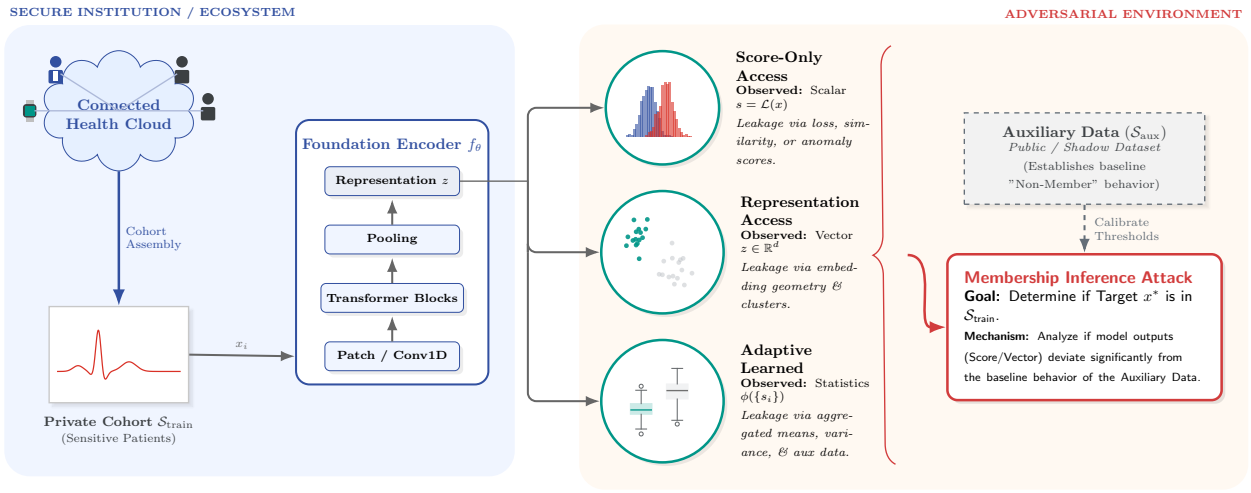


Figure 7.2: Participation privacy threat model in a connected-health ecosystem. A private patient cohort ($\mathcal{S}_{\text{train}}$) is aggregated to train a foundation ECG encoder (f_θ), which is subsequently deployed through a utility interface. An adversary probes the deployed encoder via observable outputs (e.g., scalar scores or aggregated statistics) and auxiliary public data ($\mathcal{S}_{\text{aux}}$) to infer whether a target subject contributed to pretraining.

differences motivate implementation-grounded evaluation tailored to ECG SSL pretraining.

Positioning of this work. Prior biosignal privacy work has largely studied *data-level* threats such as re-identification and linkage on shared records, whereas we focus on an underexplored risk: *model-level participation leakage* from foundation ECG encoders reused across tasks and institutions. This reuse creates new attack surfaces where an adversary can infer whether a specific individual contributed to pretraining by probing a deployed encoder or its representations, even without raw waveforms. We provide the first systematic study of membership inference for ECG foundation encoders, formalizing a connected-health threat model. Our results reveal heterogeneous and access-dependent leakage that can degrade trust and discourage data sharing in connected health.

7.3 Threat Model and Problem Definition

Figure 7.2 illustrates the connected-health threat surface we study. A foundation ECG encoder pretrained on private patient cohorts is deployed as a reusable component for downstream diagnosis, monitoring, or analytics. An adversary interacts with this deployed encoder through its exposed interface and attempts to infer whether a target individual or cohort participated in the pretraining process, resulting in a participation privacy leak.

Pretraining corpus and participation. Let a pretraining corpus consist of a set of subjects $\mathcal{S} = \{s_i\}$, where each subject s contributes a set of ECG windows $\{x_{s,j}\}_{j=1}^{n_s}$ generated by a fixed and deterministic preprocessing pipeline. A target encoder $f_\theta(\cdot)$ is pretrained on a subset $\mathcal{S}_{\text{train}} \subseteq \mathcal{S}$ and subsequently reused across tasks and deployment contexts. Given a query subject $s^\star$, the adversary aims to infer the participation indicator

$$y(s^\star) = \mathbb{I}[s^\star \in \mathcal{S}_{\text{train}}].$$

This captures *participation privacy*: even when raw waveforms and clinical labels are not exposed, membership in a training cohort can itself be sensitive, as it may reveal institutional affiliation, study participation, or cohort membership.

For auditing, we obtain ground-truth participation labels from training artifacts recorded by the pretraining pipeline (e.g., logged subject identifiers). This enables clean evaluation of membership inference without assuming oracle access by the attacker.

Observable attacker interface. We characterize the adversary by the *observable interface* of the deployed encoder, reflecting realistic connected-health deployments. We consider three interface types:

(i) **Score-only interface**: the deployment exposes only scalar outputs (e.g., reconstruction error, anomaly score, or quality metric);

(ii) **Aggregated-statistic interface**: repeated querying allows the attacker to aggregate scalar outputs across windows;

(iii) **Embedding-access interface**: the deployment exposes latent representations (feature vectors) for retrieval, clustering, or downstream analytics.

The third interface models increasingly common "feature API" deployments, where embeddings are exported for reuse without revealing model weights. Some deployments may additionally expose latent representations for downstream retrieval or analytics. While such representation access could further amplify participation leakage, a systematic empirical evaluation of representation-access attackers is beyond the scope of this submission and left to future work.

Subject-level inference and aggregation. ECG datasets typically contain many highly correlated windows per subject. Accordingly, membership inference is not meaningful at the single-window level. The adversary instead aggregates evidence across multiple queried windows from the same subject. Given an observable signal $u(\cdot)$ derived from the deployed interface, the adversary computes a subject-level score

$$\text{score}(s) = \text{Agg}\Big(\{u(x_{s,j})\}_{j=1}^{n_s}\Big),$$

and predicts participation by thresholding

$$\widehat{y}(s) = \mathbb{I}[\text{score}(s) \geq \tau].$$

We evaluate leakage using subject-level AUC, true positive rate at a fixed false-positive rate

(TPR@FPR), and attacker advantage ADV = TPR−FPR. To avoid optimistic test-time tuning, the threshold τ is selected on a held-out calibration split at a target FPR and evaluated on a disjoint test split.

Non-member construction and cross-dataset auditing. In many connected-health settings, foundation encoders are trained on near-exhaustive cohorts from a single institution or dataset, leaving few or no same-dataset subjects available as non-members for auditing. To reflect this deployment reality, we construct non-members by sampling subjects from auxiliary ECG datasets disjoint from the training dataset, yielding a *cross-dataset* evaluation protocol.

This protocol may conflate subject-level participation signals with dataset- or institution-specific characteristics, such as acquisition hardware, sensor noise, or preprocessing artifacts. Rather than interpreting this setting as strictly harder than same-dataset membership inference, we view it as modeling a realistic *cross-distribution participation leakage* threat. In practice, identifying whether a query originates from a contributing dataset or institution constitutes a meaningful privacy risk in multi-institutional connected-health deployments. We therefore interpret our results as auditing participation leakage under cross-dataset probing, rather than isolating pure within-dataset subject identification.

7.4 ECG Foundation Encoder Pretraining

We pretrain ECG foundation encoders with SSL and reuse them as fixed feature extractors in downstream connected-health pipelines (Section 7.3). Our encoder suite is aligned with widely adopted SSL recipes in recent ECG and general time-series representation learning, covering both contrastive invariance learning and masked reconstruction [25, 73, 151, 60]. This choice ensures that our privacy audit targets encoder designs that closely match cur-

rent practice in biosignal foundation modeling, rather than relying on a single specialized architecture or training objective.

Preprocessing and window construction. We convert raw WFDB ECG recordings into standardized windows and cache them on disk to keep training and evaluation consistent across datasets and runs. Signals are resampled to 250 Hz and segmented into 10-second windows with 5-second stride, which yields overlapping segments that preserve cardiac cycles while providing sufficient coverage across each subject's recording. Each recording is normalized independently to reduce amplitude-scale variation caused by sensor gain differences, acquisition settings, and patient-specific amplitude ranges. When multi-lead recordings are available, we preferentially select lead II and represent inputs as single-channel windows so that all encoders share the same input format. After preprocessing, each record yields a tensor of shape (N, C, L), where N is the number of extracted windows, C=1, and L=2000 samples per window. Caching makes the window set deterministic given a record, which is critical for membership auditing because it ensures that members and non-members are compared under exactly the same preprocessing pipeline.

Contrastive encoders. Contrastive pretraining learns a representation space in which two augmented views of the same input window map close together, while views from different windows remain separated. This strategy has become a standard foundation pretraining recipe [25] and has been adapted for ECG, where augmentations are designed to preserve clinically meaningful waveform structure while encouraging robustness to nuisance variability such as small amplitude shifts or temporal misalignment [73]. Given an input window x, we sample two stochastic augmentations $\tilde{x}^{(1)} = t_1(x)$ and $\tilde{x}^{(2)} = t_2(x)$. We encode them as $z^{(1)} = f_\theta(\tilde{x}^{(1)})$ and $z^{(2)} = f_\theta(\tilde{x}^{(2)})$, and optimize an InfoNCE [92] objective that increases similarity between positive pairs $(z^{(1)}, z^{(2)})$ and decreases similarity to other embeddings in

the same minibatch. With temperature τ and cosine similarity $\mathrm{sim}(\cdot,\cdot)$, the loss is

$$\mathcal{L}_{\mathrm{NCE}} = -\frac{1}{2B}\sum_{i=1}^{2B}\log\frac{\exp(\mathrm{sim}(z_i, z_{p(i)})/\tau)}{\sum_{j\neq i}\exp(\mathrm{sim}(z_i, z_j)/\tau)},$$

where B is the batch size and $p(i)$ indexes the paired view for sample i. Intuitively, minimizing this loss encourages the encoder to discard augmentation-specific noise while retaining stable morphological features that generalize across windows and subjects.

We report results for a SimCLR-style CNN encoder [25] and for TS2Vec [151], which extends contrastive learning to time series by explicitly enforcing agreement at multiple temporal resolutions [151]. TS2Vec computes representations that remain consistent not only at the raw window level, but also after temporal pooling (e.g., downsampling or max pooling), which encourages hierarchical features that capture both short-term waveform details and longer-range rhythm structure. In both cases, the output embedding z becomes a reusable feature representation, which is the primary object exposed by many modern foundation encoder pipelines.

Masked autoencoders. Masked reconstruction provides an alternative pretraining objective that learns representations by predicting missing portions of the input under random masking. This approach has become prominent in foundation modeling because it trains the model to capture global structure from partial observations, a property that translates well to downstream tasks where inputs may be noisy or incomplete [60]. We evaluate both a CNN-based masked autoencoder and a Transformer-based masked autoencoder with patch embedding and a lightweight decoder [60]. Given a masking pattern m (either over raw samples or over patches), we form a partially observed input and train the model to reconstruct

the original waveform as $\hat{x}$. We optimize mean squared error restricted to masked positions:

$$\mathcal{L}_{\text{MAE}} = \frac{1}{|m|} \sum_{\ell:\, m_\ell = 1} \|\hat{x}_\ell - x_\ell\|_2^2 ,$$

where $|m|$ is the number of masked elements. Restricting the loss to masked positions prevents the trivial solution of copying visible inputs, and forces the encoder to learn predictive structure that can fill in missing segments based on learned cardiac morphology and rhythm patterns. After pretraining, we extract embeddings $z = f_\theta(x)$ from the encoder pathway (e.g., by pooling token features in the Transformer model), and treat these embeddings as a realistic model output that may be exported or queried in deployment.

Pretraining regimes and membership labeling. To capture common reuse patterns in connected health, we pretrain each model family under the *single-dataset* pretraining regime. For each pretraining run, we define membership at the subject level, consistent with Section 7.3. The pretraining pipeline records the set of training subjects $\mathcal{S}_{\text{train}}$, and we assign ground-truth membership labels by

$$y(s) = \mathbb{I}[s \in \mathcal{S}_{\text{train}}].$$

This definition corresponds directly to the participation question we study: whether an individual contributed any ECG data to pretraining, regardless of which windows are later queried by an attacker.

7.5 Membership Inference for Participation Privacy

We audit participation privacy using membership inference attacks against pretrained ECG foundation encoders. These attacks provide a widely used framework for assessing whether

a trained model reveals information about its training participants through observable outputs [113]. Prior work introduced MIAs via shadow-model training and learned attackers in supervised settings [113], connected attack success to generalization gaps and loss-based tests [148], and demonstrated that attacks can remain effective under stronger adversaries with richer access to model outputs [88]. More recent studies further show that membership leakage can persist even when a model does not appear to overfit by standard metrics, motivating systematic privacy evaluation for modern large models and reusable encoders [19].

In the biosignal setting, a key distinction from canonical image or text models is that pretrained encoders are often deployed through restricted interfaces. Rather than exposing class probabilities or gradients, connected-health systems typically return scalar scores (e.g., reconstruction error or quality metrics) or aggregated statistics derived from repeated queries. Accordingly, we structure MIAs around the observable deployment interfaces described in Section 7.3, focusing on realistic attack surfaces in connected-health pipelines.

Window-level observables from deployed interfaces. Each attack operates through an observable interface exposed by the deployed encoder. Depending on the deployment setting, the adversary may observe either (i) scalar outputs derived from model behavior (e.g., reconstruction error, anomaly scores, or similarity-based statistics), or (ii) latent embeddings $z = f_\theta(x)$ returned by a feature API.

In scalar-only settings, we denote the per-window observable as a scalar signal $u(x)$ computed from the model outputs. In embedding-access settings, the observable is the window-level representation $z = f_\theta(x)$ itself, and membership inference exploits geometric structure in the latent space. Because ECG recordings are segmented into many highly correlated windows per subject, membership inference at the single-window level is neither realistic nor meaningful. Instead, a practical adversary queries multiple windows from the same subject and aggregates evidence across these queries.

Score-only black-box attacks. In the most restrictive attack setting, the adversary observes only a scalar score for each queried window. For masked autoencoders, the natural observable is reconstruction behavior: if the encoder has internalized the training distribution more strongly for member subjects, it may reconstruct their windows more accurately under the same masking pattern. To reduce variance from a single random mask realization, we evaluate reconstruction under K fixed masks and define a membership-aligned score as the negative masked reconstruction error,

$$g_{\mathrm{rec}}(x) = -\frac{1}{K}\sum_{k=1}^{K} \mathrm{MSE}_{m^{(k)}}\big(x, \hat{x}^{(k)}\big)\,,$$

where higher values indicate better reconstruction and thus stronger evidence of membership.

For contrastive encoders, deployed systems may not expose similarity scores explicitly, but repeated querying enables the computation of a *consistency* statistic. Specifically, we embed two independently augmented views of the same input window and compute their cosine similarity. Averaging over K augmentation draws yields

$$g_{\mathrm{con}}(x) = \frac{1}{K}\sum_{k=1}^{K} \cos\Big(\bar{f}_\theta(t_1^{(k)}(x)), \bar{f}_\theta(t_2^{(k)}(x))\Big)\,,$$

where $\bar{f}_\theta(\cdot)$ denotes ℓ_2-normalized embeddings. Intuitively, if training induces tighter invariances for member data due to repeated exposure to subject-specific morphology, member windows may exhibit higher representation consistency under the same augmentation pipeline.

Adaptive learned attackers. Beyond fixed scoring rules, MIAs can be strengthened when an adversary learns to combine multiple weak signals into a calibrated predictor [113, 88]. We model this capability by training a lightweight *subject-level* neural classifier on summary statistics derived from window-level scores. For each subject s, we construct a

feature vector $\mathbf{v}_s = [\mu_s, \sigma_s, \max_s, q_{0.9}(s)]$, where μ_s and σ_s denote the mean and standard deviation of $\{u(x_{s,j})\}_{j=1}^{n_s}$, $\max_s$ is the maximum score, and $q_{0.9}(s)$ is the 90th percentile. These statistics capture both typical score behavior and rare high-confidence windows that may amplify participation signals under aggregation. We train a two-layer MLP (4→16→1 with ReLU activation) using Adam (learning rate 10^{-3}) and binary cross-entropy loss for 200 optimization steps on an attacker-training split consisting of labeled member and non-member subjects. The sigmoid output is used as the subject-level membership score. As with score-only attacks, the decision threshold is selected on a disjoint calibration split to satisfy $\text{FPR} = 0.01$, and final performance is reported on held-out test subjects.

Embedding-access attacks. When the deployed encoder exposes latent representations $z = f_\theta(x)$, membership inference can be performed directly in embedding space. This models feature-API deployments where embeddings are reused for retrieval, clustering, or downstream learning without revealing model weights.

For each subject s, we sample up to T windows $\{x_{s,t}\}_{t=1}^{T}$ and compute window embeddings $e_{s,t} = f_\theta(x_{s,t})$. We construct a subject-level representation via mean pooling:

$$z_s = \frac{1}{T} \sum_{t=1}^{T} e_{s,t}.$$

The attacker builds a reference set $\mathcal{R}$ consisting of subject-level embeddings from member subjects in an attacker-training split. Membership scores are then computed using a k-nearest-neighbor (kNN) proximity rule:

$$\text{score}(s) = -\frac{1}{k} \sum_{z \in \text{kNN}(z_s, \mathcal{R})} \|z_s - z\|_2,$$

where kNN returns the k closest reference embeddings in Euclidean distance. Higher scores indicate stronger similarity to the member reference set and thus higher likelihood of partic-

ipation.

Window-to-subject aggregation. Membership inference is evaluated at the subject level. Depending on the observable interface, evidence is aggregated either from scalar window-level scores or from window-level embeddings.

In scalar-only settings, given a per-window observable $u(\cdot)$, evidence is aggregated across windows to compute a subject-level score:

$$\text{score}(s) = \text{Agg}\Big(\{u(x_{s,j})\}_{j=1}^{n_s}\Big).$$

We adopt top-k mean aggregation by default, which averages the k highest window scores for each subject. This choice reflects the empirical observation that membership cues may be concentrated in a subset of characteristic segments (e.g., unusually clean beats or distinctive morphology), rather than being uniformly distributed across all windows.

In embedding-access settings, we instead aggregate window embeddings $\{e_{s,j} = f_\theta(x_{s,j})\}$ into a subject-level representation via mean pooling:

$$z_s = \frac{1}{n_s} \sum_{j=1}^{n_s} e_{s,j},$$

We then apply the representation proximity rule from Section 7.5 to this subject-level embedding.

Calibration and leakage metrics. To report operationally meaningful privacy risk, we calibrate the decision threshold at a fixed false-positive rate (FPR) on a held-out calibration split and evaluate on a disjoint test split,

$$\widehat{y}(s) = \mathbb{I}[\text{score}(s) \geq \tau].$$

We report subject-level AUCto summarize separability without committing to a specific threshold, as well as TPRat a fixed FPR (denoted TPR@FPR) to measure attack success under controlled false-alarm budgets. Finally, we report attacker advantage ADV = TPR − FPR, which quantifies the excess success of the attacker over random guessing at the chosen operating point and is commonly used in membership auditing [148, 19].

Scope of empirical evaluation. While our threat model acknowledges that some deployments may expose learned representations, in this submission we evaluate score-only, learned subject-level, and embedding-access attackers. These interfaces collectively span realistic connected-health deployments, ranging from minimal scalar exposure to reusable feature APIs. These interfaces are directly supported by our current evaluation artifacts and reflect common connected-health deployments where only scalar outputs or aggregated statistics are accessible to external parties.

7.6 Experimental Setup

7.6.1 Datasets and preprocessing

We conduct experiments on five publicly available ECG datasets from PhysioNet. They are `butqdb` [89], `chfdb` [11], `ltdb` [56], `mitdb` [85], and `SHAREE` [83]. These datasets span diverse cardiac conditions, cohort sizes, and recording durations, and are commonly used in ECG representation learning and clinical benchmarking.

All datasets are converted into a unified window-based representation using a single deterministic preprocessing pipeline to ensure that member and non-member subjects are compared under identical transformations. Raw ECG signals are resampled to 250 Hz and segmented into 10-second windows with 5-second stride. Each window is represented as a single-channel

segment, preferentially selecting lead II when multi-lead recordings are available. We apply per-record z-normalization and remove non-finite values during preprocessing. The resulting window corpus is cached on disk and reused across all training and auditing runs.

Subject identities are defined consistently with the training pipeline. For BUTQDB, records sharing the same prefix before the first underscore are treated as belonging to the same subject, while for all other datasets each record corresponds to a unique subject. Table 7.1 summarizes the number of records, unique subjects, and extracted windows after preprocessing in the current workspace.

7.6.2 Foundation encoder pretraining

We audit four self-supervised ECG foundation encoders implemented in our codebase: SimCLR CNN, TS2Vec, MAE CNN, and MAE Transformer. All encoders are pretrained using Adam with batch size 256. We enable automatic mixed precision and apply gradient clipping with threshold 1.0 for training stability.

For contrastive encoders (SimCLR-CNN and TS2Vec), we use a temperature of 0.2. For masked autoencoders, we adopt a patch size of 50 samples and a mask ratio of 0.5. Single-dataset encoders are trained for 30 epochs. During training, we record the set of training subjects and save them as `train_ids.json`, which provides ground-truth membership labels for auditing.

7.6.3 Membership inference evaluation

We evaluate membership inference using a subject-centric auditing protocol. Because ECG datasets contain many correlated windows per subject, attacks operate on window-level outputs but are evaluated at the subject level through window-to-subject aggregation.

Attacker models and aggregation. For each subject, we sample up to $w = 2000$ windows and aggregate window-level scores using a top-k mean strategy with $k = 50$, which averages the highest-scoring windows per subject. All attacks are calibrated at a fixed false positive rate of $\text{FPR} = 0.01$ and evaluated on disjoint test subjects. We report subject-level AUC, TPR@0.01, and attacker advantage $\text{ADV} = \text{TPR} - \text{FPR}$. For learned attacks, subjects are split into attacker-train, calibration, and test subsets. The MLP is trained on attacker-train subjects with labeled membership indicators, thresholds are calibrated on the calibration subset, and results are reported on disjoint test subjects.

Score-only and learned attacks. We evaluate two attacker models aligned with realistic deployment interfaces. Score-only attacks observe only a scalar output per queried window. For MAE-based encoders, the score is the negative masked reconstruction error, averaged over eight fixed random masks. For SimCLR-CNN and TS2Vec, the score is an augmentation-consistency statistic based on cosine similarity between two independently augmented views. The learned attacker trains a lightweight subject-level classifier on summary statistics derived from window scores, including the mean, standard deviation, maximum, and upper quantile.

Embedding-space attacker. For embedding-access attacks, we construct subject-level embeddings by mean-pooling window embeddings and apply a k-nearest-neighbor proximity rule with $k = 5$. Reference embeddings are drawn from member subjects in the attacker-training split. Thresholds are calibrated at $\text{FPR} = 0.01$ on a held-out calibration subset, and final performance is reported on disjoint test subjects.

Non-member construction. In many connected-health settings, foundation encoders are trained on a near-exhaustive cohort from a given dataset, leaving few or no same-dataset subjects available as non-members. To maintain a meaningful membership inference task under such conditions, we construct non-members by sampling subjects from a mixed auxiliary

pool across ECG datasets disjoint from the training dataset. We interpret this mixed-dataset protocol as a cross-dataset robustness setting that removes trivial dataset-specific cues and measures participation leakage that persists across data sources. Unless otherwise stated, we use a balanced member-to-non-member ratio ($r = 1$) and a fixed random seed of 42.

7.7 Results

We report a systematic audit of participation privacy in self-supervised ECG foundation encoders under *single-dataset pretraining* with evaluation in a *mixed-dataset non-member* setting. Non-members are drawn from datasets disjoint from the training distribution, which models a realistic deployment risk: a party observing model behavior attempts to infer whether a specific cohort (e.g., patients from one hospital or device pipeline) contributed to pretraining. This protocol simultaneously stresses subject-level invariances (stable morphology and rhythm patterns) and dataset-level acquisition signatures, and therefore probes the kinds of population-level membership claims that arise when encoders are shared across institutions.

All metrics are calibrated at a strict false-positive rate FPR $= 0.01$ and reported as `[AUC|TPR@FPR|ADV]`. We treat AUC as a global separability measure, but emphasize TPR@0.01 and ADV as deployable indicators of tail risk under conservative auditing. Tables 7.2 and 7.3 summarize learned and score-only attacks. Table 7.4 introduces a kNN adversary with embedding access. The two figures provide complementary high-level views: Figure 7.4 visualizes dataset-level heterogeneity in learned-attack AUC across model families, and Figure 7.3 isolates the *value of learning* by plotting $\Delta\text{AUC} = \text{AUC}_{\text{learned}} - \text{AUC}_{\text{score}}$ per dataset–model pair.

7.7.1 Cohort Structure Drives Scalar Leakage

A dominant empirical pattern is that leakage is strongly regime-dependent rather than a uniform consequence of self-supervised learning. This is already apparent in Figure 7.4, where learned-attack AUC varies from near chance to near perfect depending on dataset and encoder family. The wide spread within each model column indicates that dataset properties (cohort size, redundancy, acquisition homogeneity) can dominate architectural effects, and that reporting a single "privacy number" for an encoder family is misleading without specifying the cohort regime.

The most severe scalar leakage occurs in small, homogeneous cohorts. On BUTQDB, learned attacks on SimCLR-CNN reach `[0.931|0.833|0.823]` (Table 7.2), and score-only remains comparably strong at `[0.912|0.800|0.790]` (Table 7.3). MAE-CNN is even more exposed in the scalar interface, achieving `[0.944|0.667|0.657]` in score-only mode, which is the largest score-only AUC observed across our single-dataset settings. CHFDB exhibits a similar vulnerability profile: MAE-CNN attains `[0.800|0.375|0.365]` (learned) and `[0.760|0.600|0.590]` (score-only), while MAE-Transformer shows high score-only tail risk with `[0.720|0.625|0.615]`. These results indicate that, in small cohorts, objective-aligned scalar signals (contrastive consistency or reconstruction error) often suffice to detect participation at strict calibration; learned aggregation is not required for strong separability.

In contrast, leakage attenuates on more diverse datasets and the operational gap between AUC and TPR@0.01 becomes more pronounced. MITDB shows moderate learned AUC across families (SimCLR-CNN `[0.585|0.301|0.291]`, TS2Vec `[0.594|0.260|0.250]`), but the score-only interface is much weaker for several objectives (TS2Vec `[0.647|0.000|0.000]`, MAE-CNN `[0.563|0.063|0.053]`). SHAREE exhibits a similar pattern: learned attacks reach `[0.667|0.221|0.211]` (MAE-CNN) and `[0.627|0.254|0.244]` (TS2Vec), whereas score-only for MAE-CNN collapses to `[0.490|0.000|0.000]`. Figure 7.4 captures this at-

tenuation as a compression of AUC values toward the mid-range relative to the extreme BUTQDB outliers. Empirically, increasing diversity appears to "diffuse" subject-specific stability so that only a small fraction of subjects remain separable in the extreme tail, which sharply limits TPR@0.01 even when AUC remains above chance.

7.7.2 The Value of Learning: When Learned Attackers Help (and When They Hurt)

Figure 7.3 disentangles whether training a learned attacker provides a systematic advantage over scalar thresholding. The heatmap makes clear that learned attacks are not uniformly stronger: many cells are negative, indicating that an objective-aligned scalar alone can be the worst-case interface. This observation is important operationally, because it implies that "limiting access to embeddings" or "disallowing attacker training" is insufficient as a blanket mitigation if the exposed scalar already concentrates the membership signal.

BUTQDB illustrates a score-dominated regime where learning can underperform. MAE-CNN drops from score-only AUC 0.944 to learned AUC 0.722 ($\Delta\text{AUC} = -0.222$), and TS2Vec declines from 0.778 to 0.639 ($\Delta\text{AUC} = -0.139$). In this regime, the membership signal is highly concentrated: reconstruction error (for MAE) or consistency (for contrastive) yields a sharp tail separation that a simple rule can exploit. Training a learned attacker on limited cohort diversity can instead blur this clean signal by mixing in additional window statistics that do not generalize across subjects, reducing global ranking performance even if the attacker has more capacity.

Conversely, Figure 7.3 also reveals settings where learning provides clear benefits, consistent with a diffuse-signal regime. On SHAREE, MAE-CNN improves from 0.490 (score-only) to 0.667 (learned), yielding a large positive ΔAUC, and TS2Vec increases from 0.556 to 0.627. On LTDB, MAE-CNN rises from 0.533 (score-only) to 0.596 (learned), and MAE-

Transformer decreases from 0.574 to 0.556 in AUC (a negative Δ), while SimCLR-CNN remains essentially unchanged in AUC (0.594 learned vs 0.594 score-only). The consistent takeaway is that learning helps primarily when the scalar mean is insufficient and membership cues are spread across higher-order structure (variance, stability, extremes) that must be aggregated across many windows. Figure 7.3 is therefore best interpreted not as "learning is better," but as a diagnostic of whether membership signal is concentrated into a single objective-aligned statistic or distributed across multiple summary dimensions.

7.7.3 Embedding-Attack Results: Latent Geometry Can Be the Dominant Leakage Channel

Table 7.4 adds an embedding-access adversary and qualitatively changes the threat landscape, especially for contrastive encoders. In the smallest cohorts, embedding leakage saturates: on BUTQDB and CHFDB, both SimCLR and TS2Vec achieve `[1.000|1.000|0.990]`, implying perfect subject-level detection at FPR $= 0.01$ using a simple kNN probe. This result indicates that contrastive pretraining can induce a latent geometry where member representations cluster in a way that is extremely separable from disjoint-dataset non-members, even when the attacker does not rely on scalar objective outputs. Practically, this means that releasing embeddings (or enabling downstream access to intermediate representations) may be dramatically more dangerous than exposing only logits or scores, and that the risk is not merely "incremental" over scalar leakage.

The embedding interface also clarifies an important asymmetry between contrastive and masked reconstruction objectives. While contrastive models often leak *more* through embeddings than through score-only outputs, MAE models frequently exhibit the opposite ordering. For instance, on BUTQDB, MAE-CNN embedding leakage is `[0.750|0.571|0.561]`, which is substantially below its score-only reconstruction leakage `[0.944|0.667|0.657]`; similarly

on CHFDB, MAE-CNN drops from [0.760|0.600|0.590] (score-only) to [0.710|0.429|0.419] (embedding). On LTDB, MAE embeddings are particularly well-behaved, with both MAE-CNN and MAE-Transformer yielding [0.520|0.000|0.000] and [0.510|0.000|0.000]. These patterns suggest that, for masked autoencoders, reconstruction error can amplify membership signal beyond what is directly encoded in latent neighborhood structure, whereas for contrastive methods the geometry itself can be an identity-like signature.

7.7.4 Calibration and Tail Risk at Low False-Positive Rates

Across all interfaces, strict calibration exposes tail behavior that AUC alone can obscure. Several configurations maintain moderately above-chance AUC but yield negligible TPR@0.01, indicating that separability exists in the bulk ordering but collapses when constrained to very low false positives. A clear example is TS2Vec on LTDB under learned attacks: [0.522|0.000|0.000], and under score-only: [0.556|0.000|0.000]. In contrast, other settings exhibit moderate AUC yet high tail risk, such as SimCLR on LTDB in the embedding interface ([0.750|0.667|0.657]). Operationally, this means that privacy auditing must include calibrated tail metrics: a system may appear only modestly separable globally while still enabling high-confidence membership claims for a non-trivial fraction of subjects.

These tail effects are amplified by two structural factors in our setup. First, subject-level aggregation over many windows creates an outlier-driven vulnerability: a small number of highly distinctive windows can dominate top-k summaries even if most windows overlap. Second, small cohorts with high within-subject redundancy increase the probability that some windows fall into a highly separable region (e.g., extremely low reconstruction error for a subset of beats), producing heavy-tailed member distributions. This combination explains why BUTQDB and CHFDB show exceptionally high TPR@0.01 across multiple attack types, while larger and more heterogeneous datasets compress tail separation even when average

AUC remains above chance.

7.7.5 Summary of Empirical Principles

Taken together, the tables and figures support four consistent principles. First, participation leakage is dominated by *regime heterogeneity*: Figure 7.4 shows that dataset structure can shift learned AUC from near chance to near perfect even within the same encoder family. Second, scalar interfaces are not inherently safe: Table 7.3 demonstrates that objective-aligned scores can already yield high TPR@0.01 and ADV in small cohorts. Third, the benefit of adaptive learning is conditional: Figure 7.3 shows that learning helps primarily when membership signal is diffuse, but can underperform when a single scalar statistic already concentrates separation. Fourth, embedding exposure is a distinct and sometimes dominant risk channel: Table 7.4 reveals extreme leakage for contrastive encoders, including perfect detection in small cohorts, while MAE embeddings are often less exposed than their reconstruction-based scalars.

7.8 Discussion and Limitations

Our results show that participation privacy in ECG foundation encoders depends strongly on dataset structure, training objective, and deployment interface. At the same time, several limitations of our evaluation protocol and broader practical considerations warrant discussion.

Cross-dataset auditing and confounding. Our non-member construction draws subjects from datasets disjoint from the training corpus, yielding a cross-dataset evaluation setting. This protocol may conflate subject-level membership signals with characteristics

specific to a dataset or institution, such as acquisition hardware, voltage scaling, or sensor noise. Accordingly, high attack performance does not imply pure subject identification. We adopt this protocol intentionally to reflect realistic connected-health deployments, where determining whether a query originates from a contributing institution or cohort is itself a meaningful participation risk. Nevertheless, future work should evaluate harmonized multi-site cohorts with aligned preprocessing pipelines to disentangle dataset-level artifacts from true subject-level memorization.

Preprocessing effects and signal artifacts. All datasets are resampled to a common sampling rate to enable unified pretraining. Resampling from heterogeneous original rates may introduce interpolation artifacts that are consistent within a dataset and detectable by neural encoders, potentially contributing to cross-dataset separability. A more systematic study of preprocessing sensitivity—including alternative resampling strategies and harmonized signal pipelines would clarify the extent to which observed leakage reflects physiological invariances versus signal-processing artifacts.

Broader attack surfaces and systematic evaluation. Our study focuses on score-only and subject-level learned attackers aligned with practical deployment interfaces. However, participation leakage likely spans a broader attack surface. Future work should systematically evaluate representation-access attackers, adaptive query strategies that optimize window selection, and transfer-based attacks through fine-tuned downstream models. Connected-health systems may also introduce temporal dynamics, where model updates or cohort expansion enable longitudinal membership inference. Establishing standardized evaluation protocols for biosignal MIAs—analogous to emerging benchmarks in NLP and vision—would improve comparability and rigor across studies.

Toward practical defenses. Our findings suggest that defenses must be objective-aware and deployment-aware. Potential directions include differentially private pretraining, calibrated noise injection, or regularization techniques that limit extreme-tail separability in reconstruction or consistency scores. However, privacy mechanisms must preserve clinically meaningful morphology and diagnostic utility. Developing principled trade-offs between representation quality and participation privacy remains an open challenge, particularly for reusable foundation encoders integrated into clinical infrastructure.

7.9 Conclusion

We presented a systematic audit of participation privacy in self-supervised ECG foundation encoders. By evaluating score-only and learned membership inference attacks under a deployment-aligned, subject-centric protocol, we showed that commonly used SSL objectives can leak training participation even in the absence of labels or explicit identity features. Importantly, we found that strong leakage is concentrated in small or institution-specific cohorts, and that restricting access to scalar scores is insufficient to guarantee privacy when those scores align closely with the pretraining objective.

As reusable biosignal encoders become integral components of connected-health infrastructure, our results underscore the need for rigorous, context-aware privacy auditing prior to deployment. Future work will extend this analysis to representation-access attackers, harmonized multi-site cohorts, and principled defenses that balance utility and participation privacy.

Table 7.1: **Diverse real-world ECG corpora used in this study.** We report (i) *intrinsic dataset properties* (cohort, acquisition setting, original sampling rate, available leads, and typical record duration) from the official PhysioNet releases, and (ii) *the effective scale used in our workspace* after our deterministic preprocessing. **Preprocessing (shared across all datasets):** we resample all signals to 250 Hz, extract 10-second windows with 5-second stride, select a single ECG channel (prefer lead II when available; otherwise use an available ECG lead), remove non-finite values, and apply per-record z-normalization before caching windows to disk. These datasets jointly cover heterogeneous acquisition devices, cohort sizes, and recording durations, reflecting practical connected-health conditions rather than a single curated benchmark.

Dataset	**Cohort / acquisition (from PhysioNet)**	**Orig. f_s**	**Leads**	**Typical length**	**#Used windows**
BUTQDB [89]	Single-lead long-term recordings collected for ECG quality assessment; Holter-style ambulatory signals (wearable / mobile setting).	1000 Hz	1	$\geq$24 h	344,660
CHFDB [11]	Long-term ECG from congestive heart failure patients; continuous ambulatory recordings.	250 Hz	2	$\sim$20–24 h	215,177
LTDB [56]	Long-term ST/ischemia-related ECG with prolonged ambulatory monitoring; annotated ST episodes in long recordings.	250 Hz	1–2	$\sim$21–24 h	106,157
MIT-BIH Arrhythmia [85]	Clinical arrhythmia benchmark; annotated two-lead excerpts widely used for rhythm/beat analysis.	360 Hz	2	30 min	20,017
SHAREE [83]	Ambulatory recordings from a hypertensive cohort with Holter-style monitoring; multi-signal release includes ECG channels.	128 Hz	2 ECG (+aux)	24 h	2,296,925

Table 7.2: **Learned membership inference.** Results represent [AUC|TPR@0.01|Adv] evaluated under a **mixed-dataset non-member setting**. Bold indicates the highest leakage for each dataset.

Dataset	SimCLR CNN	TS2Vec	MAE CNN	MAE Transformer
BUTQDB	[0.931|0.833 |0.823]	[0.639|0.375 |0.365]	[0.722|0.333 |0.323]	[0.528|0.200 |0.190]
CHFDB	[0.620|0.333 |0.323]	[0.560|0.333 |0.323]	**[0.800|0.375 |0.365]**	[0.734|0.333 |0.323]
LTDB	[0.594|0.333 |0.323]	[0.522|0.000 |0.000]	**[0.596|0.500 |0.490]**	[0.556|0.333 |0.323]
MITDB	[0.585|**0.301 |0.291]**	**[0.594**|0.260 |0.250]	[0.592|0.210 |0.200]	[0.591|0.196 |0.186]
SHAREE	[0.647|0.231 |0.221]	[0.627|**0.254 |0.244]**	**[0.667**|0.221 |0.211]	[0.588|0.117 |0.107]

Note: w=2000 windows/subject, r=1 non-member ratio. Calibration at FPR=0.01.

Table 7.3: **Score-only membership inference.** Comparison of leakage using only scalar outputs. Small non-zero artifacts in TPR are corrected to 0.000, and negative advantage scores are clipped to 0.000. Bold indicates the highest leakage for each dataset.

Dataset	SimCLR CNN	TS2Vec	MAE CNN	MAE Transformer
BUTQDB	[0.912|**0.800 |0.790]**	[0.778|0.333 |0.323]	**[0.944**|0.667 |0.657]	[0.556|0.143 |0.133]
CHFDB	[0.580|0.250 |0.240]	[0.600|0.000 |0.000]	**[0.760**|0.600 |0.590]	[0.720|**0.625 |0.615]**
LTDB	**[0.594**|0.250 |0.240]	[0.556|0.000 |0.000]	[0.533|0.000 |0.000]	[0.574|**0.333 |0.323]**
MITDB	[0.639|0.221 |0.211]	**[0.647**|0.000 |0.000]	[0.563|0.063 |0.053]	[0.606|**0.320 |0.310]**
SHAREE	**[0.637|0.137 |0.127]**	[0.556|0.000 |0.000]	[0.490|0.000 |0.000]	[0.599|0.000 |0.000]

Note: Same protocol as Table 7.2.

Table 7.4: **Embedding-access membership inference.** Results ([AUC|TPR@0.01|Adv]) for an adversary probing the latent representation space using kNN. Contrastive models (SimCLR, TS2Vec) show extreme leakage in embedding space, while MAE models often leak *less* through embeddings than through reconstruction error (compare to Table 7.3).

Dataset	**SimCLR CNN**	**TS2Vec**	**MAE CNN**	**MAE Transformer**
BUTQDB	[1.000|1.000|0.990]	[1.000|1.000|0.990]	[0.750|0.571|0.561]	[0.610|0.286|0.276]
CHFDB	[1.000|1.000|0.990]	[1.000|1.000|0.990]	[0.710|0.429|0.419]	[0.550|0.143|0.133]
LTDB	[0.750|0.667|0.657]	[0.680|0.333|0.323]	[0.520|0.000|0.000]	[0.510|0.000|0.000]
MITDB	[0.680|0.348|0.338]	[0.660|0.304|0.294]	[0.550|0.174|0.164]	[0.540|0.130|0.120]
SHAREE	[0.690|0.261|0.251]	[0.680|0.275|0.265]	[0.580|0.145|0.135]	[0.520|0.058|0.048]

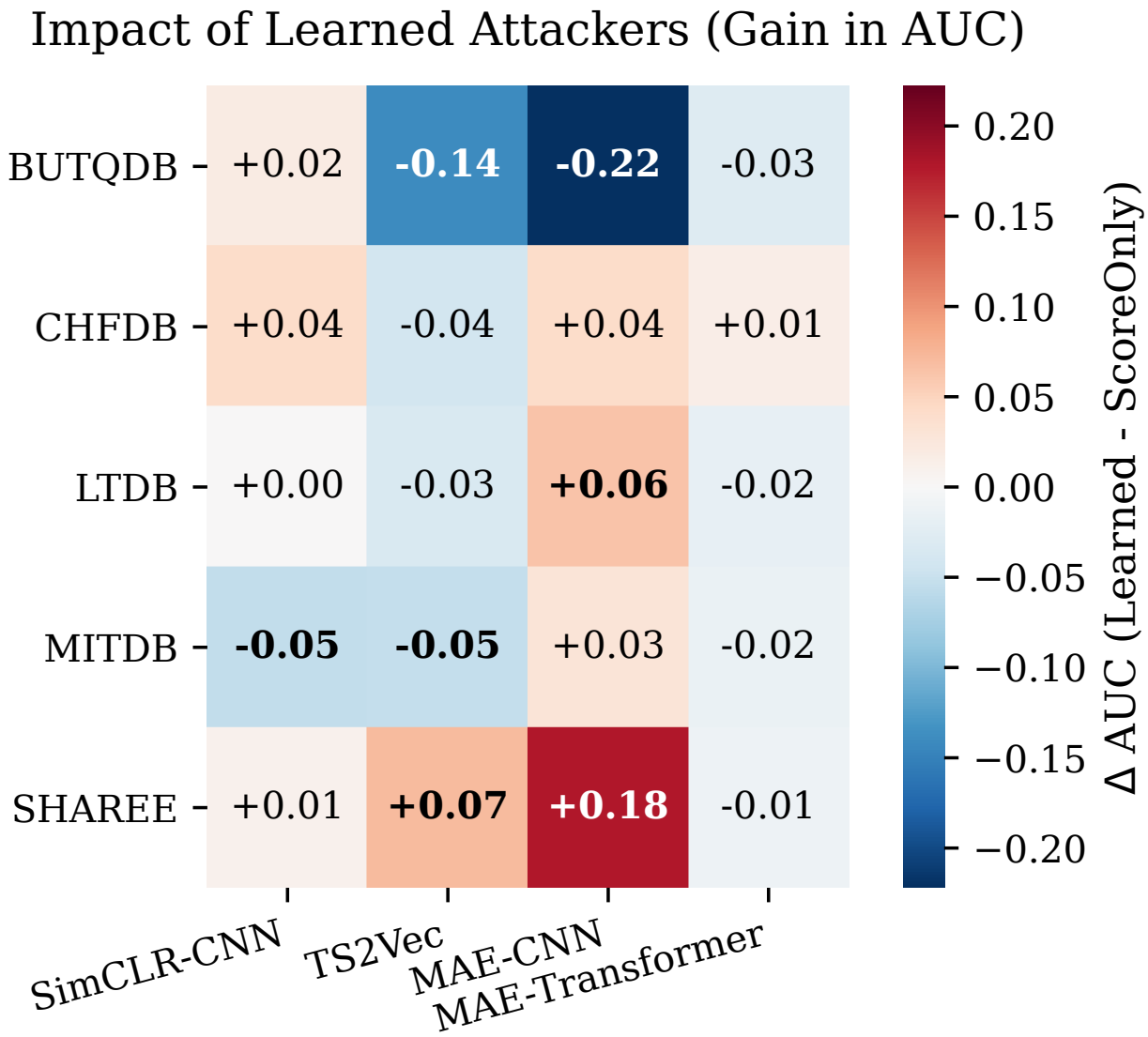


Figure 7.3: Impact of learned attackers over score-only attacks in single-dataset pretraining. Each cell reports $\Delta\text{AUC} = \text{AUC}_{\text{learned}} - \text{AUC}_{\text{score}}$. Positive values indicate that training a learned attacker improves separability; negative values indicate that an objective-aligned score-only rule is stronger.

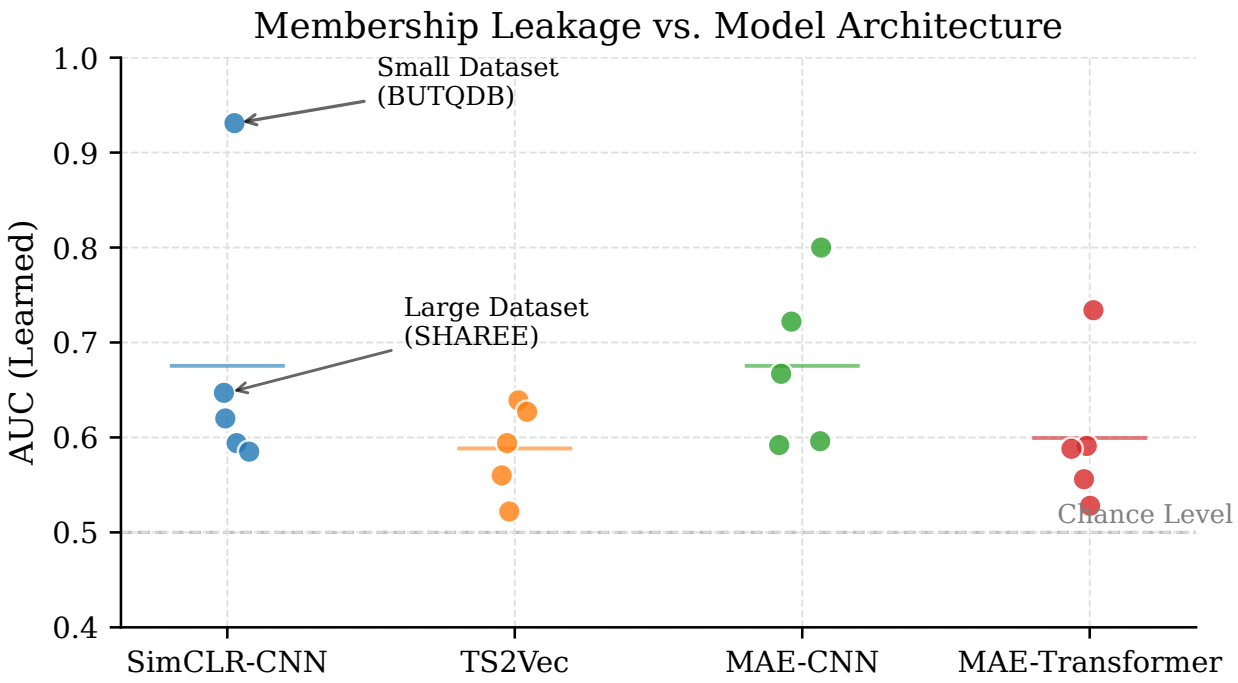


Figure 7.4: Learned-attack AUC across datasets and encoder families for single-dataset pretraining. The dashed line indicates chance-level AUC of 0.5. Leakage varies substantially across objectives and datasets, and the most severe cases occur in small-cohort settings.

Chapter 8

Summary and Future Directions

8.1 Summary of the Dissertation

This dissertation investigated how healthcare AI can adapt to users and data without treating privacy or participant burden as afterthoughts. Its central argument is that trustworthy adaptation is not a single model property. It is a system property that must be designed, tested, and maintained from user interaction and data collection through model deployment.

Chapter 2 began with the interaction between an everyday monitoring system and the person who supplies its labels. The context-aware active reinforcement-learning framework modeled whether an ecological momentary assessment would be informative and timely using prediction uncertainty, response history, and contextual information. Offline results demonstrated substantial reductions in the number of queries required to reach a target performance, while the online study deployed real-time triggering and evaluated context features and personalization. This chapter established that adaptation begins before model training: a monitoring system must learn when and how to ask for information without exhausting the user.

Chapter 3 then addressed how to learn securely from the sparse, distributed observations produced in everyday monitoring. Differential private federated transfer learning combined three complementary ideas: transfer learning addressed limited local data, federated learning reduced the need to centralize sensitive observations, and differential privacy limited the information exposed through shared updates. In a real-world stress-monitoring case study, the framework improved accuracy and recall relative to learning from scarce local data alone while providing an explicit privacy mechanism. Together, the first two studies connected adaptive data acquisition with privacy-preserving adaptation.

The remaining chapters examined what such mechanisms do not automatically protect. Chapter 4 showed that ECG data retain biometric identity even after explicit identifiers are removed. Transparent classifiers and SHAP explanations connected re-identification performance to interpretable PQRST features, demonstrating that privacy risk is rooted in the physiology of the signal rather than solely in a particular model architecture.

Chapter 5 then moved from handcrafted features to learned representations. TransECG showed that transformer attention can exploit identity-related information distributed across long ECG sequences and can localize that information to meaningful cardiac components. The performance gain over feature-based, convolutional, and recurrent baselines illustrated a recurring dual-use property of healthcare AI: richer representations improve analytical power, but the same representations can make sensitive structure more accessible.

Chapter 6 converted this re-identification capability into a practical adversarial setting. The linkage framework removed the unrealistic assumption that every target participant is known to the attacker. Even with partial knowledge and unknown participants, a two-stage attack could link ECG records across datasets at substantial accuracy. This result demonstrated why anonymization and closed-set evaluation provide insufficient evidence of privacy.

Finally, Chapter 7 moved the attack surface from released data to released models. The

membership-inference audit showed that scalar outputs and latent embeddings from ECG foundation encoders can disclose training participation. Leakage varied with the pretraining objective, cohort structure, attacker interface, and operating point. Particularly for small or institution-specific cohorts, restricting access to raw signals or labels was not enough to guarantee participation privacy.

8.2 Cross-Cutting Findings

Six findings connect the chapters.

First, data quality and user burden are coupled. A monitoring system that requests too many labels, or asks at inappropriate moments, can reduce participation and degrade the very data needed for personalization. Second, decentralization is not equivalent to privacy. Federated learning changes where computation occurs, but formal protection and empirical auditing remain necessary. Third, physiological utility and identity leakage arise from overlapping signal structure. A transformation that removes all individual variation may also remove clinically useful information, so defenses must be evaluated against both privacy and downstream utility. Fourth, representation learning amplifies this tension: models that learn richer temporal and morphological dependencies may improve health inference while creating stronger identity or membership signals. Fifth, attack realism matters. Open-set recognition, heterogeneous sources, repeated subject-level queries, and low-false-positive operating points can change conclusions that would be drawn from aggregate accuracy or AUC alone. Sixth, interpretability is useful for both science and security. Mapping query decisions, leakage, summary statistics, signal regions, and embedding geometry helps convert model behavior into actionable knowledge for designers and data stewards.

These findings suggest a progression from privacy by mechanism to privacy by evidence. A

formal mechanism defines a protection goal, but a trustworthy healthcare system must also demonstrate through context-aware testing that the remaining interfaces do not violate the expectations of patients, study participants, institutions, or developers.

8.3 Future Research Directions

8.3.1 Human-Centered Adaptive Monitoring

Future monitoring systems should jointly optimize label utility, participant burden, fairness, and privacy rather than treating EMA scheduling as an isolated prediction problem. Reinforcement-learning policies could incorporate individual interruption preferences, uncertainty about delayed or missing responses, and explicit limits on daily or cumulative burden. Longitudinal evaluation should measure not only model accuracy but also adherence, response latency, perceived intrusiveness, and whether scheduling quality differs across demographic or clinical groups. Fairness evaluation must also extend to synthetic medical data used to supplement scarce observations: subgroup imbalance can persist or be amplified during generation, motivating augmentation frameworks that explicitly improve demographic representation before synthetic-data training [103]. Privacy-preserving on-device context modeling would further reduce the need to transmit detailed behavioral data merely to decide when to ask a question.

8.3.2 End-to-End Privacy Accounting

Future systems should connect privacy guarantees across stages that are usually studied separately. Context collection, label querying, differentially private training, model fine-tuning, repeated API queries, embedding release, and downstream personalization can each

reveal information or consume a privacy budget. An end-to-end accounting framework should describe how these operations compose over time and across institutions. Such a framework would make it possible to compare a system's stated guarantee with the leakage observed under realistic identity, linkage, and membership attacks.

8.3.3 Utility-Aware Representation Protection

The ECG studies show that identity and clinical information can occupy overlapping regions of the signal and latent space. Future defenses should therefore optimize a multi-objective criterion: retain diagnostic and monitoring utility while suppressing identity and participation cues. Promising directions include adversarially learned identity-invariant representations, differentially private self-supervised objectives, feature-selective perturbation, representation bottlenecks, and calibrated embedding-release policies. Evaluation should include multiple clinical tasks so that apparent privacy gains are not achieved by indiscriminately degrading the signal.

8.3.4 Longitudinal and Cross-Device Evaluation

Biometric characteristics are neither perfectly constant nor completely unstable. Aging, disease progression, medication, stress, sensor placement, sampling rate, and device hardware can change ECG morphology. Larger longitudinal studies are needed to determine when identity cues persist, when attacks transfer across devices or institutions, and whether defenses remain effective under these shifts. These studies should emphasize participant-level separation and demographic diversity to avoid overstating generalization.

8.3.5 Causal and Clinically Grounded Explanations

SHAP values and transformer attention identify associations between model decisions and ECG regions, but they do not by themselves establish causal mechanisms. Future work should combine attribution with controlled perturbations, counterfactual signal generation, electrophysiological priors, and clinical review. This would clarify whether a highlighted region truly carries stable identity information or merely correlates with dataset-specific acquisition artifacts. Clinically grounded explanations can also help determine which transformations are safe for diagnosis but effective for privacy.

8.3.6 Adaptive Auditing and Deployment Controls

Privacy risk varies across cohorts, objectives, and interfaces, so a single benchmark threshold is unlikely to be sufficient. Deployment-aware auditing should select attacks and operating points based on the consequences of false accusations and missed leakage. Systems can then adapt controls accordingly: limit repeated queries, round or randomize scalar outputs, restrict embeddings, monitor suspicious access patterns, or require stronger authentication for high-risk interfaces. Audits should be repeated after model updates because fine-tuning and distribution shift may create new leakage channels.

8.3.7 Governance for Physiological Data and Models

Technical defenses should be paired with governance that recognizes physiological signals as both health data and potential biometrics. Consent processes should explain risks from cross-dataset linkage and model participation, not only direct disclosure. Data-use agreements should define whether embeddings and pretrained encoders may be redistributed, and model documentation should report the populations used in training, the interfaces exposed, and the

privacy audits performed. These measures can align the incentives of researchers, healthcare institutions, technology providers, and participants.

8.4 Concluding Perspective

Healthcare AI will increasingly learn across people, devices, institutions, and tasks. Its value will depend on adaptation, but its legitimacy will depend on whether that adaptation respects the people whose effort and data make it possible. The work in this dissertation traces a path from context-aware, burden-conscious label acquisition to privacy-preserving distributed learning and then to progressively deeper audits of signals, representations, cross-dataset identity, and training participation. The resulting lesson is not that useful health data should cease to be shared or that expressive models should cease to be developed. It is that every increase in interaction, reuse, and representational power should be accompanied by corresponding evidence of human benefit and privacy protection. Designing adaptive mechanisms and testing their residual risks together offers a practical foundation for secure, responsive, and trustworthy healthcare AI.

Bibliography

[1] A. Alahmadi et al. A privacy-preserved iomt-based mental stress detection framework with federated learning. *The Journal of Supercomputing*, pages 1–20, 2023.

[2] H. Alikhani, A. Kanduri, P. Liljeberg, A. M. Rahmani, and N. Dutt. Dynafuse: Dynamic fusion for resource efficient multi-modal machine learning inference. *IEEE Embedded Systems Letters*, 2023.

[3] H. Alikhani, Z. Wang, A. Kanduri, P. Lilieberg, A. M. Rahmani, and N. Dutt. Seal: Sensing efficient active learning on wearables through context-awareness. In *2024 Design, Automation & Test in Europe Conference & Exhibition (DATE)*, pages 1–2. IEEE, 2024.

[4] H. Alikhani, Z. Wang, A. Kanduri, P. Liljeberg, A. M. Rahmani, and N. Dutt. Ea^ 2: Energy efficient adaptive active learning for smart wearables. In *Proceedings of the 29th ACM/IEEE international symposium on low power electronics and design*, pages 1–6, 2024.

[5] J. Allen. Photoplethysmography and its application in clinical physiological measurement. *Physiological measurement*, 28(3):R1, 2007.

[6] American Institute of Stress. The american institute of stress, 2021.

[7] S. A. H. Aqajari, Z. Wang, A. Tazarv, S. Labbaf, S. Jafarlou, B. Nguyen, N. Dutt, M. Levorato, and A. M. Rahmani. Enhancing performance and user engagement in everyday stress monitoring: A context-aware active reinforcement learning approach. Accepted for publication in ACM Transactions on Computing for Healthcare (HEALTH), 2026.

[8] E. A. Ashley and J. Niebauer. *Cardiology explained*. Remedica, 2004.

[9] AWARE Framework. Aware framework, 2023.

[10] Bagdasaryan et al. How to backdoor federated learning. In *International conference on artificial intelligence and statistics*, pages 2938–2948. PMLR, 2020.

[11] D. S. Baim, W. S. Colucci, E. S. Monrad, H. S. Smith, R. F. Wright, A. Lanoue, D. F. Gauthier, B. J. Ransil, W. Grossman, and E. Braunwald. Survival of patients with severe congestive heart failure treated with oral milrinone. *Journal of the American College of Cardiology*, 7(3):661–670, 1986.

[12] S. L. Battalio et al. Sense2stop: a micro-randomized trial using wearable sensors to optimize a just-in-time-adaptive stress management intervention for smoking relapse prevention. *Contemporary Clinical Trials*, 109:106534, 2021.

[13] L. Biel, O. Pettersson, L. Philipson, and P. Wide. Ecg analysis: a new approach in human identification. *IEEE transactions on instrumentation and measurement*, 50(3):808–812, 2001.

[14] L. Breiman. Random forests. *Machine learning*, 45:5–32, 2001.

[15] S. Bucci et al. The digital revolution and its impact on mental health care. *Psychology and Psychotherapy: Theory, Research and Practice*, 92(2):277–297, 2019.

[16] L. E. Burke, S. Shiffman, E. Music, M. A. Styn, A. Kriska, A. Smailagic, D. Siewiorek, L. J. Ewing, E. Chasens, B. French, et al. Ecological momentary assessment in behavioral research: addressing technological and human participant challenges. *Journal of medical Internet research*, 19(3):e77, 2017.

[17] Y. S. Can et al. Real-life stress level monitoring using smart bands in the light of contextual information. *IEEE Sensors Journal*, 20(15):8721–8730, 2020.

[18] Y. S. Can et al. Privacy-preserving federated deep learning for wearable iot-based biomedical monitoring. *ACM Transactions on Internet Technology (TOIT)*, 21(1):1–17, 2021.

[19] N. Carlini, S. Chien, M. Nasr, S. Song, A. Terzis, and F. Tramèr. Membership inference attacks from first principles. In *2022 IEEE Symposium on Security and Privacy (SP)*, pages 1897–1914. IEEE, 2022.

[20] D. Castaneda et al. A review on wearable photoplethysmography sensors and their potential future applications in health care. *International journal of biosensors & bioelectronics*, 4(4):195, 2018.

[21] P. Chandra, K. Jussi, P.-M. Mari, and A.-S. Katriina. Sex differences in heart: from basics to clinics. *European Journal of Medical Research (Web)*, 27(1):1–17, 2022.

[22] P. H. Charlton et al. Assessing mental stress from the photoplethysmogram: a numerical study. *Physiological measurement*, 39(5):054001, 2018.

[23] C. Che, P. Zhang, M. Zhu, Y. Qu, and B. Jin. Constrained transformer network for ecg signal processing and arrhythmia classification. *BMC Medical Informatics and Decision Making*, 21(1):184, 2021.

[24] T. Chen et al. Xgboost: extreme gradient boosting. *R package version 0.4-2*, 1(4):1–4, 2015.

[25] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning (ICML)*, pages 1597–1607. PMLR, 2020.

[26] Y. Chen et al. Fedhealth: A federated transfer learning framework for wearable healthcare. *IEEE Intelligent Systems*, 35(4):83–93, 2020.

[27] M. Cheng, X. Diao, Z. Zhou, Y. Cui, W. Liu, and S. Cheng. Toward short-term glucose prediction solely based on cgm time series, 2024.

[28] M. Cheng et al. Saic: Integration of speech anonymization and identity classification, 2023.

[29] M. Cheng et al. Vetrass: Vehicle trajectory similarity search through graph modeling and representation learning. *arXiv preprint arXiv:2404.08021*, 2024.

[30] M. Cheng, Z. Zhou, B. Zhang, Z. Wang, J. Gan, Z. Ren, W. Feng, Y. Lyu, H. Zhang, and X. Diao. Efflex: Efficient and flexible pipeline for spatio-temporal trajectory graph modeling and representation learning. *arXiv preprint arXiv:2404.12400*, 2024.

[31] K. K. Coelho, E. T. Tristão, M. Nogueira, A. B. Vieira, and J. A. Nacif. Multimodal biometric authentication method by federated learning. *Biomedical Signal Processing and Control*, 85:105022, 2023.

[32] S. Cohen, T. Kamarck, R. Mermelstein, et al. Perceived stress scale. *Measuring stress: A guide for health and social scientists*, 10(2):1–2, 1994.

[33] A. Datta, T. Bhattacharyya, E. Khatibi, A. Seth, Z. Wang, S. R. Mousavi, A. M. Rahmani, F. Firouzi, and K. Chakrabarty. React: Reinforcement learning-based adaptive ecg anonymization and privacy threat mitigation. In *2025 IEEE International Conference on Omni-layer Intelligent Systems (COINS)*, pages 1–8. IEEE, 2025.

[34] N. K. Dixit et al. Managing the scarcity of monitoring data through machine learning for human behavior in mental health care. In *Applications of Deep Learning and Big IoT on Personalized Healthcare Services*, pages 147–175. IGI Global, 2020.

[35] G. Dogan et al. Stress detection using experience sampling: A systematic mapping study. *International Journal of Environmental Research and Public Health*, 19(9):5693, 2022.

[36] A. Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[37] A. Dubey, A. Jauhri, A. Pandey, et al. The llama 3 herd of models, 2024. arXiv:2407.21783.

[38] Z. Ebrahimi, M. Loni, M. Daneshtalab, and A. Gharehbaghi. A review on deep learning methods for ecg arrhythmia classification. *Expert Systems with Applications: X*, 7:100033, 2020.

[39] D. M. et al. NeuroKit2: A python toolbox for neurophysiological signal processing. *Behavior Research Methods*, 53(4):1689–1696, Feb 2021.

[40] G. R. et al. Explainable deep learning: A field guide for the uninitiated. *J. Artif. Intell. Res.*, 73, 2022.

[41] K. N. P. et al. Ecg biometric recognition without fiducial detection. In *Proc. Biometrics Symp. Special Sess. Res. Biometric Consortium Conf.* IEEE, 2006.

[42] M. I. et al. Ecg biometric authentication: A comparative analysis. *IEEE Access*, 8, 2020.

[43] P. L. et al. Explainable ai: A review of machine learning interpretability methods. *Entropy*, 23(1), 2020.

[44] R. V. et al. *ECG diagnosis in clinical practice*. Springer, 2009.

[45] S. H. et al. Aspects of privacy for electronic health records. *Int. J. Med. Inform.*, 80(2), 2011.

[46] S. K. B. et al. A survey on ecg analysis. *Biomed. Signal Process. Control*, 43:216–235, 2018.

[47] U. Y. et al. Evaluation of ppg biometrics for authentication in different states. In *Proc. Int. Conf. Biometrics (ICB)*. IEEE, 2018.

[48] X. Y. et al. Zebra: Deeply integrating system-level provenance search and tracking for efficient attack investigation. *arXiv preprint arXiv:2211.05403*, 2022.

[49] J. Fahrenberg, M. Myrtek, K. Pawlik, and M. Perrez. Ambulatory assessment-monitoring behavior in daily life settings. *European Journal of Psychological Assessment*, 23(4):206–213, 2007.

[50] M. Fredrikson et al. Model inversion attacks that exploit confidence information and basic countermeasures. In *Proceedings of the 22nd ACM SIGSAC conference on computer and communications security*, pages 1322–1333, 2015.

[51] E. Garcia-Ceja et al. Mental health monitoring with multimodal sensing and machine learning: A survey. *Pervasive and Mobile Computing*, 51:1–26, 2018.

[52] A. Ghazarian, J. Zheng, H. El-Askary, H. Chu, G. Fu, and C. Rakovski. Increased risks of re-identification for patients posed by deep learning-based ecg identification algorithms. In *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 1969–1975. IEEE, 2021.

[53] A. Ghazarian, J. Zheng, D. Struppa, and C. Rakovski. Assessing the reidentification risks posed by deep learning algorithms applied to ecg data. *IEEE Access*, 10:68711–68723, 2022.

[54] G. Giannakakis, D. Grigoriadis, K. Giannakaki, O. Simantiraki, A. Roniotis, and M. Tsiknakis. Review on psychological stress detection using biosignals. *IEEE Transactions on Affective Computing*, 13(1):440–460, 2019.

[55] M. Gjoreski, M. Luštrek, M. Gams, and H. Gjoreski. Monitoring stress with a wrist device using context. *Journal of biomedical informatics*, 73:159–170, 2017.

[56] A. L. Goldberger, L. A. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley. Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals. *circulation*, 101(23):e215–e220, 2000.

[57] S. D. Greenwald. *The development and analysis of a ventricular fibrillation detector*. PhD thesis, Massachusetts Institute of Technology, 1986.

[58] A. Gu, K. Goel, and C. Ré. Efficiently modeling long sequences with structured state spaces. *arXiv preprint arXiv:2111.00396*, 2021.

[59] H. J. Han et al. Objective stress monitoring based on wearable sensors in everyday settings. *Journal of Medical Engineering & Technology*, 44(4):177–189, 2020.

[60] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16000–16009, 2022.

[61] M. A. Hearst et al. Support vector machines. *IEEE Intelligent Systems and their applications*, 13(4):18–28, 1998.

[62] T. Hester et al. Deep q-learning from demonstrations. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.1, 2018.

[63] T. Hlaing, T. DiMino, P. R. Kowey, and G.-X. Yan. Ecg repolarization waves: their genesis and clinical implications. *Annals of noninvasive electrocardiology*, 10(2):211–223, 2005.

[64] T. H. Holmes and R. H. Rahe. The social readjustment rating scale. *Journal of psychosomatic research*, 1967.

[65] H. Hu, Z. Qiao, M. Cheng, Z. Liu, and H. Wang. Dasgil: Domain adaptation for semantic and geometric-aware image-based localization. *IEEE Transactions on Image Processing*, 30:1342–1353, 2020.

[66] J. W. Hurst. Naming of the waves in the ecg, with a brief account of their genesis. *Circulation*, 98(18), 1998.

[67] A. Kanduri, S. Shahhosseini, E. K. Naeini, H. Alikhani, P. Liljeberg, N. Dutt, and A. M. Rahmani. *Edge-Centric Optimization of Multi-modal ML-Driven eHealth Applications*, pages 95–125. Springer Nature Switzerland, Cham, 2024.

[68] B. J. Kenny and K. N. Brown. Ecg t wave. In *StatPearls [Internet]*. StatPearls Publishing, 2022.

[69] E. Khatibi, Z. Wang, and A. M. Rahmani. Cdf-rag: Causal dynamic feedback for adaptive retrieval-augmented generation. *arXiv preprint arXiv:2504.12560*, 2025.

[70] E. Khatibi, Z. Wang, A. Sharma, K. Chakrabarty, S. R. Moosavi, F. Firouzi, and A. Rahmani. Care-ecg: Causal agent-based reasoning for explainable and counterfactual ecg interpretation. *arXiv preprint arXiv:2604.10420*, 2026.

[71] A. N. Kho, J. P. Cashy, K. L. Jackson, A. R. Pah, S. Goel, J. Boehnke, J. E. Humphries, S. D. Kominers, B. N. Hota, S. A. Sims, et al. Design and implementation of a privacy preserving electronic health record linkage tool in chicago. *Journal of the American Medical Informatics Association*, 22(5):1072–1080, 2015.

[72] B.-H. Kim and J.-Y. Pyun. Ecg identification for personal authentication using lstm-based deep recurrent neural networks. *Sensors*, 20(11):3069, 2020.

[73] D. Kiyasseh, T. Zhu, and D. A. Clifton. Clocs: Contrastive learning of cardiac signals across space, time, and patients. In *International Conference on Machine Learning*, pages 5606–5615. PMLR, 2021.

[74] S. B. Kotsiantis et al. Supervised machine learning: A review of classification techniques. *Emerging artificial intelligence applications in computer engineering*, 160(1):3–24, 2007.

[75] R. D. Labati, E. Muñoz, V. Piuri, R. Sassi, and F. Scotti. Deep-ecg: Convolutional neural networks for ecg biometric recognition. *Pattern Recognition Letters*, 126:78–85, 2019.

[76] S. Labs. Optimizing current consumption in bluetooth low energy devices, 2024. Available at: `https://docs.silabs.com/bluetooth/5.0/general/system-and-performance/optimizing-current-consumption-in-bluetooth-low-energy-devices`.

[77] F. Larradet et al. Toward emotion recognition from physiological signals in the wild: approaching the methodological issues in real-life data collection. *Frontiers in psychology*, 11:1111, 2020.

[78] J. Lee, D.-H. Jang, S. Park, and S. Cho. A low-power photoplethysmogram-based heart rate sensor using heartbeat locked loop. *IEEE Transactions on Biomedical Circuits and Systems*, 12(6):1220–1229, 2018.

[79] Z. C. Lipton. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 16(3), 2018.

[80] Mayo Clinic. Mayo clinic, 2023.

[81] K. McKeen, S. Masood, A. Toma, B. Rubin, and B. Wang. Ecg-fm: An open electrocardiogram foundation model. *JAMIA open*, 8(5):ooaf122, 2025.

[82] B. McMahan et al. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR, 2017.

[83] P. Melillo, R. Izzo, A. Orrico, P. Scala, M. Attanasio, M. Mirra, N. De Luca, and L. Pecchia. Automatic prediction of cardiovascular and cerebrovascular events using heart rate variability analysis. *PloS one*, 10(3):e0118504, 2015.

[84] F. Monitillo, M. Leone, C. Rizzo, A. Passantino, and M. Iacoviello. Ventricular repolarization measures for arrhythmic risk stratification. *World journal of cardiology*, 8(1):57, 2016.

[85] G. B. Moody and R. G. Mark. The impact of the mit-bih arrhythmia database. *IEEE engineering in medicine and biology magazine*, 20(3):45–50, 2001.

[86] K. Mundnich et al. Tiles-2018, a longitudinal physiologic and behavioral data set of hospital workers. *Scientific Data*, 7(1):354, 2020.

[87] M. Nasr et al. Comprehensive privacy analysis of deep learning. In *Proceedings of the 2019 IEEE Symposium on Security and Privacy (SP)*, pages 1–15, 2018.

[88] M. Nasr, R. Shokri, and A. Houmansadr. Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning. In *2019 IEEE Symposium on Security and Privacy (SP)*, pages 739–753. IEEE, 2019.

[89] A. Nemcova, R. Smisek, K. Opravilová, M. Vitek, L. Smital, and L. Maršánová. Brno University of Technology ECG Quality Database (BUT QDB). *PhysioNet*, July 2020. Version 1.0.0.

[90] A. Nolin-Lapalme, R. Avram, and H. Julie. Privecg: generating private ecg for end-to-end anonymization. In *Machine Learning for Healthcare Conference*, pages 509–528. PMLR, 2023.

[91] I. Odinaka, P.-H. Lai, A. D. Kaplan, J. A. O'Sullivan, E. J. Sirevaag, and J. W. Rohrbaugh. Ecg biometric recognition: A comparative analysis. *IEEE Transactions on Information Forensics and Security*, 7(6):1812–1824, 2012.

[92] A. v. d. Oord, Y. Li, and O. Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[93] OpenAI. Gpt-4 technical report, 2023. arXiv:2303.08774.

[94] M. Pelc, Y. Khoma, and V. Khoma. Ecg signal as robust and reliable biometric marker: Datasets and algorithms comparison. *Sensors*, 19(10):2350, 2019.

[95] A. Pillai, D. Spathis, F. Kawsar, and M. Malekzadeh. PaPaGei: Open Foundation Models for Optical Physiological Signals. In *International Conference on Learning Representations (ICLR)*, 2025. arXiv preprint arXiv:2410.20542.

[96] M. Plappert. keras-rl. `https://github.com/keras-rl/keras-rl`, 2016.

[97] T. Polonelli et al. H-watch: An open, connected platform for ai-enhanced covid19 infection symptoms monitoring and contact tracing. In *2021 IEEE International Symposium on Circuits and Systems (ISCAS)*, pages 1–5, 2021.

[98] J. Powar and A. R. Beresford. Sok: Managing risks of linkage attacks on data privacy. *Proc. Privacy Enhancing Technol.*, 2023.

[99] M. Prince et al. No health without mental health. *The lancet*, 370(9590):859–877, 2007.

[100] M. M. Qirtas et al. Privacy preserving loneliness detection: A federated learning approach. In *2022 IEEE International Conference on Digital Health (ICDH)*, pages 157–162. IEEE, 2022.

[101] A. M. L. Research. Large-scale training of foundation models for wearable biosignals. Apple Machine Learning Research, 2024.

[102] M. Saha, M. A. Xu, W. Mao, S. Neupane, J. M. Rehg, and S. Kumar. Pulse-ppg: An open-source field-trained ppg foundation model for wearable applications across lab and field settings. *arXiv preprint arXiv:2502.01108*, 2025.

[103] S. Salarian, Y. Zhang, S. Padhee, and S. Parthasarathy. MedEqualizer: A framework investigating bias in synthetic medical data and mitigation via augmentation. *arXiv preprint arXiv:2511.01054*, 2025.

[104] G. Sannino et al. A mobile system for real-time context-aware monitoring of patients' health and fainting. *International journal of data mining and bioinformatics*, 10(4):407–423, 2014.

[105] A. Sano and R. W. Picard. Stress recognition using wearable sensors and mobile phones. In *2013 Humaine association conference on affective computing and intelligent interaction*, pages 671–676. IEEE, 2013.

[106] F. Sarhaddi et al. A comprehensive accuracy assessment of samsung smartwatch heart rate and heart rate variability. *PloS one*, 17(12):e0268361, 2022.

[107] Y. Sattar and L. Chhabra. Electrocardiogram. In *StatPearls [Internet]*. StatPearls Publishing, 2023.

[108] J. Schläpfer and H. J. Wellens. Computer-interpreted electrocardiograms: benefits and limitations. *Journal of the American College of Cardiology*, 70(9):1183–1192, 2017.

[109] P. Schmidt, A. Reiss, R. Duerichen, C. Marberger, and K. Van Laerhoven. Introducing wesad, a multimodal dataset for wearable stress and affect detection. In *Proceedings of the 20th ACM international conference on multimodal interaction*, pages 400–408, 2018.

[110] D. Seok et al. Motion artifact removal techniques for wearable eeg and ppg sensor systems. *Frontiers in Electronics*, 2:685513, 2021.

[111] B. Settles. Active learning literature survey. Technical Report 1648, University of Wisconsin–Madison, Department of Computer Sciences, 2009.

[112] Z. Sharifi-Heris et al. Phenotyping the autonomic nervous system in pregnancy using remote sensors: potential for complication prediction. *Frontiers in Physiology*, 14:1293946, 2023.

[113] R. Shokri, M. Stronati, C. Song, and V. Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*, pages 3–18. IEEE, 2017.

[114] D. Sidebotham. *Cardiothoracic critical care*. Elsevier Health Sciences, 2007.

[115] L. Sina et al. Zotcare: a flexible, personalizable, and affordable mhealth service provider. *Front. Digit. Health*, 2023.

[116] T. Song, G. Lu, and J. Yan. Emotion recognition based on physiological signals using convolution neural networks. In *Proceedings of the 2020 12th international conference on machine learning and computing*, pages 161–165, 2020.

[117] A. Steiner, A. Kolesnikov, X. Zhai, R. Wightman, J. Uszkoreit, and L. Beyer. How to train your vit? data, augmentation, and regularization in vision transformers. *arXiv preprint arXiv:2106.10270*, 2021.

[118] A. Steptoe, S. Kunz-Ebrecht, N. Owen, P. J. Feldman, G. Willemsen, C. Kirschbaum, and M. Marmot. Socioeconomic status and stress-related biological responses over the working day. *Psychosomatic medicine*, 65(3):461–470, 2003.

[119] M. Stojchevska et al. Assessing the added value of context during stress detection from wearable data. *BMC Medical Informatics and Decision Making*, 22(1):268, 2022.

[120] B. Surawicz and T. Knilans. *Chou's electrocardiography in clinical practice: adult and pediatric*. Elsevier Health Sciences, 2008.

[121] B. Surawicz and S. R. Parikh. Differences between ventricular repolarization in men and women: description, mechanism and implications. *Annals of Noninvasive Electrocardiology*, 8(4):333–340, 2003.

[122] B. Suruliraj et al. Federated learning framework for mobile sensing apps in mental health. In *2022 IEEE 10th International Conference on Serious Games and Applications for Health (SeGAH)*, pages 1–7. IEEE, 2022.

[123] L. M. Swift, M. Burke, D. Guerrelli, M. Reilly, M. Ramadan, D. McCullough, T. Prudencio, C. Mulvany, A. Chaluvadi, R. Jaimes, et al. Age-dependent changes in electrophysiology and calcium handling: implications for pediatric cardiac research. *American Journal of Physiology-Heart and Circulatory Physiology*, 318(2):H354–H365, 2020.

[124] J. Taelman, S. Vandeput, A. Spaepen, and S. Van Huffel. Influence of mental stress on heart rate and heart rate variability. In *4th European Conference of the International Federation for Medical and Biological Engineering: ECIFMBE 2008 23–27 November 2008 Antwerp, Belgium*, pages 1366–1369. Springer, 2009.

[125] A. Tazarv et al. Personalized stress monitoring using wearable sensors in everyday settings. In *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 7332–7335. IEEE, 2021.

[126] A. Tazarv et al. Active reinforcement learning for personalized stress monitoring in everyday settings. In *2023 IEEE/ACM Conference on Connected Health (CHASE)*, pages 44–55, Los Alamitos, CA, USA, jun 2023. IEEE Computer Society.

[127] M. P. Turakhia, M. Desai, R. A. Harrington, et al. Rationale and design of a large-scale, app-based study to identify cardiac arrhythmias using a smartwatch: The apple heart study. *American Heart Journal*, 207:66–75, 2019.

[128] P. Van Gent et al. Heartpy: A novel heart rate algorithm for the analysis of noisy signals. *Transportation research part F: traffic psychology and behaviour*, 66:368–378, 2019.

[129] C. Van Mieghem, M. Sabbe, and D. Knockaert. The clinical value of the ecg in noncardiac conditions. *Chest*, 125(4):1561–1576, 2004.

[130] G. Vashisht et al. A study on the tizen operating system. *International Journal of Computer Trends and Technology*, 12(1):14–15, 2014.

[131] P. Wagner, N. Strodthoff, R.-D. Bousseljot, D. Kreiseler, F. I. Lunze, W. Samek, and T. Schaeffter. Ptb-xl, a large publicly available electrocardiography dataset. *Scientific Data*, 7(1):154, 2020.

[132] C. Wang and J. Guo. A data-driven framework for learners' cognitive load detection using ecg-ppg physiological feature fusion and xgboost classification. *Procedia computer science*, 147:338–348, 2019.

[133] W. Wang et al. Social sensing: assessing social functioning of patients living with schizophrenia using mobile phone sensing. In *Proceedings of the 2020 CHI conference on human factors in computing systems*, pages 1–15, 2020.

[134] Z. Wang, A. Kanduri, S. A. H. Aqajari, S. Jafarlou, S. R. Mousavi, P. Liljeberg, S. Malik, and A. M. Rahmani. ECG unveiled: Analysis of client re-identification risks in real-world ECG datasets. In *2024 IEEE 20th International Conference on Body Sensor Networks (BSN)*, pages 1–4, 2024.

[135] Z. Wang, E. Khatibi, F. Firouzi, S. R. Mousavi, K. Chakrabarty, and A. M. Rahmani. Linkage attacks expose identity risks in public ECG data sharing. In *2025 47th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 1–7, 2025.

[136] Z. Wang, E. Khatibi, K. Kazemi, I. Azimi, S. Mousavi, S. Malik, and A. M. Rahmani. TransECG: Interpreting biometric identity leakage in ECG signals via vision transformers. In *22nd IEEE-EMBS International Conference on Body Sensor Networks (BSN 2026)*. IEEE, 2026.

[137] Z. Wang, E. Khatibi, and A. M. Rahmani. Medcot-rag: Causal chain-of-thought rag for medical question answering. In *2025 IEEE 21st International Conference on Body Sensor Networks (BSN)*, pages 1–4. IEEE, 2025.

[138] Z. Wang, E. Khatibi, A. Sharma, K. Chakrabarty, S. R. Moosavi, F. Firouzi, and A. Rahmani. Membership inference attacks expose participation privacy in ECG foundation encoders. *Smart Health*, 41:100680, 2026.

[139] Z. Wang, H. Li, D. Huang, H.-S. Kim, C.-W. Shin, and A. M. Rahmani. Healthq: Unveiling questioning capabilities of llm chains in healthcare conversations. *Smart Health*, page 100570, 2025.

[140] Z. Wang, N. Luo, and P. Zhou. Guardhealth: Blockchain empowered secure data management and graph convolutional network enabled anomaly detection in smart healthcare. *Journal of Parallel and Distributed Computing*, 142:1–12, 2020.

[141] Z. Wang, Z. Yang, I. Azimi, and A. M. Rahmani. Differential private federated transfer learning for mental health monitoring in everyday settings: A case study on stress detection. In *2024 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 5402–5406, 2024.

[142] X. Wei, S. Yohannan, and J. R. Richards. Physiology, cardiac repolarization dispersion and reserve. In *StatPearls [Internet]*. StatPearls Publishing, Treasure Island, FL, 2023. Updated April 17, 2023.

[143] X. Xu et al. Fedmood: Federated learning on mobile health data for mood detection. *arXiv preprint arXiv:2102.09342*, 2021.

[144] A. Yang, B. Yang, B. Hui, et al. Qwen2 technical report, 2024. arXiv:2407.10671.

[145] Z. Yang et al. Loneliness forecasting using multi-modal wearable and mobile sensing in everyday settings. In *2023 IEEE 19th International Conference on Body Sensor Networks (BSN)*, pages 1–4. IEEE, 2023.

[146] Z. Yang et al. Chatdiet: Empowering personalized nutrition-oriented food recommender chatbots through an llm-augmented framework. *Smart Health*, page 100465, 2024.

[147] Y. Yao, Z. Wang, and P. Zhou. Privacy-preserving and energy efficient task offloading for collaborative mobile computing in iot: An admm approach. *Computers & Security*, 96:101886, 2020.

[148] S. Yeom, I. Giacomelli, M. Fredrikson, and S. Jha. Privacy risk in machine learning: Analyzing the connection to overfitting. In *2018 IEEE 31st Computer Security Foundations Symposium (CSF)*, pages 268–282. IEEE, 2018.

[149] H. Yu et al. Semi-supervised learning and data augmentation in wearable-based momentary stress detection in the wild. *arXiv preprint arXiv:2202.12935*, 2022.

[150] H. Yu and A. Sano. Passive sensor data based future mood, health, and stress prediction: User adaptation using deep learning. In *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 5884–5887, 2020.

[151] Z. Yue, Y. Wang, J. Duan, T. Yang, C. Huang, Y. Tong, and B. Xu. Ts2vec: Towards universal representation of time series. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 8980–8987, 2022.

[152] Y. Zhang, S. Padhee, P. T. Yuhas, C. J. Roberts, and S. Parthasarathy. Adaptive temporal mixture of experts for predicting stiffness metrics from the ocular response analyzer and identifying keratoconus. *American Journal of Ophthalmology*, 286:196–210, 2026.

[153] Y. Zhang, M. Reidy, S. Padhee, N. Doble, D. Mutti, and S. Parthasarathy. Automated montaging of ultrasound scans for determination of ocular globe shape. *Research Square*, 2026.

[154] Y. Zhang, C. J. Roberts, S. Padhee, N. Doble, S. Parthasarathy, and P. T. Yuhas. Machine learning detection of keratoconus through ocular response analyzer waveform parameters. *Investigative Ophthalmology & Visual Science*, 66(8):3378, 2025.

[155] M. Zhou, Y. Zhang, A. Karimi Monsefi, S. S. Choi, N. Doble, S. Parthasarathy, and R. Ramnath. Reducing manual labeling requirements and improved retinal ganglion cell identification in 3D AO-OCT volumes using semi-supervised learning. *Biomedical Optics Express*, 15(8):4540–4556, 2024.

[156] M. Zhou, Y. Zhang, E. Kirkendall, A. Karimi Monsefi, M. Wolfe, K. A. Chitkara, S. S. Choi, N. Doble, S. Parthasarathy, and R. Ramnath. ISOSNet: A unified framework for cone photoreceptor detection and inner segment and outer segment length measurement from AO-OCT b-scans. *Biomedical Optics Express*, 16(8):3237–3254, 2025.

[157] Z. Zhou et al. Crossgp: Cross-day glucose prediction excluding physiological information. *arXiv preprint arXiv:2404.10901*, 2024.

[158] Z. Zhou et al. Glumarker: A novel predictive modeling of glycemic control through digital biomarkers, 2024.