Biomedical Signal Processing and Control 112 (2026) 108502

Contents lists available at ScienceDirect

Biomedical Signal Processing and Control

journal homepage: www.elsevier.com/locate/bspc

Securing deep learning models with differential privacy for cardiovascular disease prediction

Zoher Orabe *, Antti Vasankari, Tapio Pahikkala, Matti Kaisti, Antti Airola

Department of Computing, University of Turku, Turku, FI-20014, Finland

ARTICLE INFO

Dataset link: https://github.com/UTU-Health-Research/dl-ecg-dp-classifier/

Keywords:
Cardiovascular diseases
Deep learning
Differential privacy
Electrocardiography

ABSTRACT

This study investigates how differential privacy (DP) can enhance data confidentiality in deep learning models for predicting cardiovascular diseases (CVDs) using electrocardiography (ECG) data collected from various hospitals. We evaluated the privacy–utility trade-off by analyzing model performance under different privacy budgets (ϵ) across different model architectures, including the high-capacity ResNet with squeeze-and-excitation (ResNet-SE), transformer-based model, and two simple baselines: logistic regression (LR) and multilayer perceptrons (MLP). The original ResNet-SE model, with 8.81 million parameters, showed substantial performance degradation under DP with macro- and micro-average AUCs decreasing from 0.90 and 0.92 to 0.79 and 0.82 at $\epsilon = 10$. By reducing the model size by 98.4% to 142,934 parameters, we achieved a better balance between accuracy and privacy, with macro- and micro-average AUCs of 0.87 and 0.89, only 0.03 lower than its non-private performance. The transformer-based model showed weaker robustness to DP, with a macro- and micro-average AUCs dropping from 0.88 and 0.91 to 0.64 and 0.73, while LR and MLP baselines trained on ECG handcrafted features achieved low performance even without privacy. The effect of training with DP varied across classes, having only minimal impact on the four largest classes (AUC reduction ≤ 0.01), but more substantial performance decreases were observed for many of the smaller classes (e.g. 0.10 drop for a condition with a 1.19% class size, and a drop of 0.28 for condition with class size of 3.10%). Overall, our study demonstrates the positive effect of reducing model complexity for improving privacy-utility trade-off for predicting CVDs.

1. Introduction

Cardiovascular diseases (CVDs) are a complex group of disorders involving the heart and blood vessels, often influenced by genetic, behavioral, and environmental factors. They represent a major cause of global morbidity and mortality [1]. Therefore, it is crucial to mitigate the impact of CVD by developing advanced and accurate diagnostic tools that can assist in identifying CVD at its earliest stages. Early detection helps to improve patient outcomes and reduces the healthcare burden, as timely intervention can significantly reduce associated morbidity and mortality.

Machine learning (ML) models have shown great potential in detecting CVDs and enhancing clinical decision-making because of their capabilities to analyze complex datasets and identify patterns that may not be detectable by traditional methods. Recent studies have demonstrated the efficacy of ML models, particularly deep learning models, in predicting CVD with high accuracy [2,3].

However, applying machine learning models to healthcare problems requires dealing with sensitive patient data. One of the primary concerns is the potential leakage of private information, which can occur through various attacks on ML models. For instance, model inversion attacks have been shown to successfully reconstruct sensitive input data, such as medical images, from trained ML models [4].

This raises critical privacy concerns, as unauthorized access to patient data can lead to severe consequences, including breaches of confidentiality and loss of patient trust. Multi-hospital datasets allow enhancing model robustness and improving the detection of rare conditions that may be underrepresented in individual datasets [5,6]. The data sources examined in our study, hosted by PhysioNet/CinC Challenge [7], are significantly imbalanced, with significant differences in label distributions across hospitals. For example, certain cardiovascular conditions are underrepresented in one hospital's dataset but sufficiently represented in another as outlined in [8]. However, integrating

* Corresponding author.
E-mail addresses: zoher.orabe@utu.fi (Z. Orabe), antti.s.vasankari@utu.fi (A. Vasankari), aatapa@utu.fi (T. Pahikkala), mkaist@utu.fi (M. Kaisti), ajairo@utu.fi (A. Airola).

https://doi.org/10.1016/j.bspc.2025.108502
Received 19 February 2025; Received in revised form 30 June 2025; Accepted 2 August 2025
Available online 1 September 2025

data from multiple hospitals helps mitigate this imbalance, improving class representation across the board.

This aggregation also enables more reliable evaluation of a model's ability to generalize to previously unseen distributions from different clinical settings. However, when models trained on such aggregated datasets are shared beyond the participating institutions whether through public release or deployment in external clinical environments there is a risk of inadvertently exposing sensitive information, such as patient personal data, through model inversion or membership inference attacks [4]. To address this concern, our study incorporates differential privacy (DP) during model training to ensure the confidentiality of patient information in the resulting models. Generally, DP has emerged as a promising solution to the privacy challenges faced by machine learning models [9].

Legally, DP has been suggested as a potential technology for achieving compliance with data protection legislation, such as GDPR [10]. it provides a robust mathematical framework to protect individual privacy within datasets by ensuring that the presence or absence of any individual in the dataset cannot be inferred. This is achieved by incorporating controlled noise into the learning process, such as by bounding and adding noise to gradients during model training [11]. it allows hospitals to share aggregated or statistical data without compromising individual privacy, facilitating collaboration.

While DP effectively reduces the risk of sensitive information leakage, it introduces a trade-off between privacy and model accuracy. The noise added to maintain privacy can degrade model performance, which poses a particular challenge in critical applications such as healthcare, where accuracy is paramount. Since patients' lives are at stake, significant loss of accuracy in clinical prediction models is unacceptable. Our study demonstrates the impact of applying differential privacy on CVDs individual classes, especially the rare ones. However, DP can disproportionally affect the performance of underrepresented classes due to added noise [12]. Recent studies [13,14] have demonstrated that applying differential privacy to ECG deep learning models with strict privacy budgets results in significant performance degradation, making the models impractical for clinical deployment. This highlights the challenge of developing DP methods for ECG models that enforce strong privacy protections while maintaining an acceptable level of utility.

In this study, we present empirical evidence on the effectiveness of DP with deep learning for training clinical prediction models. We specifically apply DP to a state-of-the-art ECG classifier based on the ResNet-SE architecture [15], which was originally developed for the PhysioNet/CinC Challenge and ranked among the top contenders. This model classifies various CVDs by analyzing 12-lead ECG signals and clinical information. While deep learning models are often highly overparameterized to capture complex representations in ECG classification tasks, Ponomareva et al. [16] have shown that deep learning models with fewer parameters require less noise to be injected during training under DP conditions. In our study, we investigate whether reducing the size of the ResNet-SE trainable parameters improves the accuracy-privacy trade-off when classifying CVDs. To assess the generalizability of our findings across different model architectures, we also compared the impact of DP on the state-of-the-art transformer-based ECG model [17], and two simpler baselines: logistic regression and MLP, which are trained on ECG handcrafted features.

The primary research questions we address are:

RQ1: How does enforcing varying levels of DP affect the accuracy of a state-of-the-art ECG deep learning classifiers?

RQ2: Can a more favorable accuracy-privacy trade-off be achieved by reducing model architecture size?

RQ3: What is the impact of DP on the performance of individual classes, especially underrepresented classes?

In Section 2, we review related works. Section 3 details the datasets, machine learning methodologies, and DP approach used. Section 4 presents the experimental results, followed by Section 5, which provides discussion and conclusion. Finally, Section 6 outlines the study's limitations and potential directions for future work.

2. Related work

Abadi et al. [11] proposed DP-SGD, a differentially private stochastic gradient descent algorithm for training deep learning models. This approach clips gradients, adds noise, and employs a privacy accountant to monitor privacy loss during training. Further studies [18,19] extended DP to other optimizers, such as Adam [20]. In our study, we adopt the Adam optimizer, as it yielded the best performance for the original ECG classifier in the absence of DP.

Bagdasaryan et al. [12] showed that DP disproportionately affects underrepresented and complex subgroups, resulting in notable accuracy reductions in tasks like sentiment analysis and image classification. Ponomareva et al. [16] investigated how architectural choices such as activation functions, optimizers, and model sizes impact the performance of DP-trained deep learning models. They concluded that no definitive theoretical or empirical guidance exists for optimizing model architecture under DP, treating architecture as a hyperparameter requiring careful tuning. They further emphasized the critical role of privacy-specific hyperparameters, including clipping norms, noise multipliers, and privacy budgets, in maximizing model utility under DP constraints.

Applying DP to clinical prediction models has gained traction due to increasing concerns over patient data privacy. DP has been successfully integrated into healthcare tasks such as cardiovascular disease prediction from electronic health records [21], genomic data analysis [22], deep survival models [23], and medical image analysis [24,25]. Notably, Yan et al. [25] explored architectural impacts on DP-trained models for medical image classification, highlighting the efficacy of low-rank adaptation in reducing the adverse effects of DP noise.

However, existing works on DP deep learning for ECG classification highlight a severe drop in performance when privacy is enforced. As summarized in Table 1, models trained without any privacy constraints achieve high performance (e.g. CNNs reach AUC of 0.80–0.88, BiLSTMs ACC of 0.93), but performance degrades sharply as the privacy budget tightens. For instance, Agrawal et al. [13] report a drop in AUC from 0.80 (No-DP) to 0.55 at $\epsilon = 5$ and further down to 0.50 at $\epsilon = 0.5$. Similarly, Zhang et al. [14] observe BiLSTM accuracy falling from 0.93 (No-DP) to 0.68 at $\epsilon = 15$, and Gil and Vejar [26] show their DP-SGD CNN maintaining only 0.72–0.73 accuracy at budgets ($\epsilon = 12$–120). Islam and Imtiaz's CNN–BiLSTM heart-rate estimator sees its mean squared error rise from 0.90 (No-DP) to 5.86 at $\epsilon = 1.87$ under strict privacy constraints [27]. These results collectively demonstrate that enforcing strong privacy levels (e.g., $\epsilon \leq 10$) often renders ECG models unsuitable for real-life applications.

In our study, we address this challenge by investigating how to balance the privacy–utility trade-off while preserving acceptable performance levels for models trained under DP on ECG data collected from multiple hospitals. A different approach to securing ECG multi-hospital data was discussed by Weimann and Conrad [28], who applied federated learning (FL) to ECG data from the CinC2021 challenge, the same dataset used in our study. They developed a ResNet-based model for multi-label ECG classification and demonstrated that FL outperforms models trained in isolation. However, the global model remains vulnerable to attacks such as membership inference, a threat that DP can help mitigate. This limitation motivated our investigation into the application of DP to protect the model against such attacks and to assess its impact on model performance, especially under high levels of privacy. Additionally, other DP-based methods have been proposed, including sanitizing individual ECG time series or generating synthetic ECGs. These approaches aim to facilitate public sharing of sensitive, sanitized ECG data while keeping the original ECG on the individual's device [29,30].

Table 1
Performance comparison of machine learning models across varying differential privacy levels in recent biomedical studies.

Study	Model	Metric	Privacy budget (ϵ)	Score	Remarks
[13]	CNN	AUC	(No-DP)	0.80	DP + FL across 7 hospitals; AUC drops with stricter privacy
			100	0.65	
			5	0.55	
			0.5	0.50	
[14]	BiLSTM	ACC	(No-DP)	0.93	Two-stage DP in FL; user-level + server-side noise
			60	0.90	
			30	0.83	
			15	0.68	
[26]	CNN	ACC	(No-DP)	0.88	DP-SGD CNN for arrhythmia detection; real-time use-case
			120	0.73	
			48	0.73	
			12	0.72	
[27]	CNN–BiLSTM	MSE	(No-DP)	0.90	Heart rate estimation from ECG + PPG; error increases under DP
			37.5	2.23	
			8.6	2.75	
			1.87	5.86	

Note: CNN = Convolutional Neural Network, LSTM = Long Short-Term Memory, FL = Federated Learning, AUC = Area Under the Curve, ACC = Accuracy, MSE = Mean Squared Error.

3. Material and methods

3.1. Differential privacy

The foundation of differential privacy was introduced in [31], which provides a framework aimed at ensuring strong privacy protections in the analysis and sharing of data. A randomized mechanism $\mathcal{M}$ is said to be (ϵ, δ)-differentially private if, for any two datasets D and D' that differ by at most one element, and for any subset of outputs $S \subseteq \text{Range}(\mathcal{M})$, the probability that $\mathcal{M}$ outputs the subset S from dataset D is not significantly different from the probability of obtaining the same output from dataset D'. This relationship can be expressed mathematically as follows.

$$\Pr[\mathcal{M}(D) \in S] \leq e^{\epsilon} \Pr[\mathcal{M}(D') \in S] + \delta \tag{1}$$

where ϵ and $\delta \leq 1$ are non-negative scalars known as the privacy parameters. Choosing the privacy budget $\epsilon \leq 1$ provides a strong privacy guarantee according to the definition of differential privacy [32]. However, values less than or equal to 1 can often lead to a significant drop in utility for larger machine learning models and may not be practical. Reasonable privacy guarantees often suggest using $\epsilon \leq 10$, based on various real-world DP applications and academic studies [11]. Even a very high ϵ can be seen as an improvement over models with no privacy protections, as it allows for the quantification of potential privacy gains. The parameter δ represents the probability of privacy leakage, allowing for a small chance of the mechanism failing to meet the (ϵ, δ) privacy guarantee. A smaller δ ensures stronger privacy guarantees by reducing the likelihood of an adversary inferring sensitive information. Common practice recommends setting $\delta \leq \frac{1}{N}$, where N is the dataset size, to minimize expected privacy breaches to well below one instance, even under worst-case scenarios [16].

In our study, we employ Rényi Differential Privacy (RDP) [33], an extension of DP that uses Rényi divergence to provide tighter bounds on the cumulative privacy loss over multiple computations. RDP offers improved composition properties, making it suitable for deep learning applications that involve many training iterations.

Applying Differential Privacy to Deep Learning: In deep learning, training consists of two main processes: forward propagation, which involves computing the loss, and backward propagation, which updates the model using gradients. In the DP-SGD framework developed by Abadi et al. [11], the raw training data is only utilized during the computation of the gradient for each individual example. Immediately after the gradient for each sample is calculated, the algorithm clips its L2 norm (to limit its influence) and adds DP noise. The remaining steps (including averaging the gradients and updating the weights) are carried out solely on these sanitized gradients. According to the DP post-processing theorem [34], any further operations (such as updates, accumulation of parameters, or forward passes to compute the loss) cannot compromise privacy. The implementation of DP-SGD consists of two key steps:

1. Gradient clipping: This step limits the influence of individual training examples on the overall gradient. By setting a maximum threshold for the L2-norm of the gradients, known as the clipping norm (C), any gradient that exceeds this threshold is clipped so that its norm equals (C). This process helps reduce the risk of data leakage, ensuring that no single example can disproportionately impact the model's updates.

2. Noise addition: After clipping, noise is added to the gradients. Noise is scaled according to the privacy parameters (ϵ) and (δ), as well as the total number of training steps and sampling rate, using the RDP accountant to track cumulative privacy loss. Adding the noise ensures that even if someone accesses the trained model, they cannot easily infer information about the individual training examples.

The steps for Differentially Private Adam (DP-Adam) are similar to those of DP-SGD. In DP-Adam, the algorithm computes the first and second moment estimates of the gradients using the noisy, clipped gradients. By updating these moment estimates based on differentially private gradients, the optimizer maintains effective learning rates while ensuring that differential privacy is preserved. All of our ECG deep learning models including ResNet-based, Transformer-based, and MLP architectures were trained using the DP-Adam optimizer with the same previous steps that include gradient clipping and noise addition.

3.2. Datasets

This section provides a comprehensive overview of the datasets used in this study, which include data from the PhysioNet/Computing in Cardiology Challenge 2021 (CinC) [35–37], as well as an additional publicly available database from Shandong Provincial Hospital (SPH) [38].

The CinC dataset comprises four distinct sources: the CPSC and CPSC-Extra datasets [39], the PTB and PTBXL datasets [40,41], the

Table 2
Number of patient samples from each data source used in our study.

Source	Number of Samples
SPH	23,274
G12EC	8827
PTB & PTBXL	21,256
CPSC & CPSC-Extra	6102
Chapman-Shaoxing & Ningbo	43,560

Table 3
ECG condition representation with their class ratios.

Condition	Code	Class ratio
Sinus Rhythm (SR)	426783006	46.55%
Sinus Bradycardia (SB)	426177001	21.16%
T Wave Abnormality (TWA)	164934002	12.10%
Sinus Tachycardia (ST)	427084000	9.66%
Atrial Flutter (AFL)	164890007	8.74%
Left Axis Deviation (LAD)	39732003	7.22%
Atrial Fibrillation (AF)	164889003	5.67%
Sinus Arrhythmia (SA)	427393009	5.18%
T Wave Inversion (TWI)	59931005	3.54%
1st Degree AV Block (1° AVB)	270492004	3.45%
Right Bundle Branch Block (RBBB)	59118001	3.41%
Premature Atrial Contraction (PAC)	284470004	3.10%
Incomplete Right Bundle Branch Block (IRBBB)	713426002	2.86%
Left Anterior Fascicular Block (LAFB)	445118002	2.27%
Low QRS Voltage (LQRSV)	251146004	1.64%
Right Axis Deviation (RAD)	47665007	1.43%
Left Bundle Branch Block (LBBB)	164909002	1.19%

Chapman-Shaoxing and Ningbo datasets [42,43], and the Georgia 12-lead ECG (G12EC) dataset [7].

These datasets contain a variety of clinical information, including patient demographics, alongside 12-lead ECG recordings from individuals diagnosed with various cardiovascular diseases. Each label encoded using standardized formats such as SNOMED CT[1] for the CinC dataset and AHA[2] for the SPH dataset. Notably, all datasets are multi-label, indicating that each patient can be diagnosed with multiple cardiovascular conditions simultaneously, However some of the labels are mutually exclusive. Most of the ECGs in the CinC dataset were sampled at 500 Hz, whereas a smaller portion from the PTB and PTB-XL datasets were sampled at 1000 Hz. The recordings varied in length from 5 s to nearly 2.5 min. In the SPH dataset, ECGs were sampled at 500 Hz with durations ranging between 10 and 60 s. Most datasets included demographic data such as age and gender, along with detailed information on lead configuration and recording settings. Given the variety of data sources, it was essential to standardize and harmonize the labels to ensure consistency for experimental analyses. Leinonen et al. [8] developed a preprocessing pipeline to process all sources, selecting 17 labels from a large label set based on criteria related to the CinC challenge. Each label was required to be present in at least four distinct data sources. They also resampled the data to 250 Hz and adjusted each ECG to a length of 4096 samples. We relied on their open-source code[3] to preprocess and prepare our datasets.

Table 2 provides a summary of the number of patient samples collected from each source, the data show significant variability in sample sizes among the sources. Furthermore, Table 3 outlines the classification of various cardiac conditions within our dataset, complete with their corresponding codes. The class ratios, presented in the last column, were calculated based on the training samples.

[1] https://www.snomed.org/
[2] https://www.heart.org/en/professional
[3] https://github.com/tuijalei/12-lead-ecg-classifier

3.3. Modeling with raw ECG signals

Residual Network with Squeeze-and-Excitation (ResNet-SE) The model architecture is illustrated in Fig. 2 originally developed by [15] for the PhysioNet/CinC Challenge, employs a ResNet architecture supplemented with squeeze-and-excitation blocks to enhance performance. The model begins with input layers for age and gender, which are processed through a fully connected layer with ReLU activation before being concatenated with features from convolutional layers. The ECG 12-lead signals are processed through a one-dimensional (1D) convolutional layer with a kernel size of 15, resulting in 64 output channels and subsequent max pooling to compact the features. It comprises four residual layers, each consisting of two BasicBlocks that include two 1D convolutional layers with a kernel size of 7, enhanced by squeeze-and-excitation blocks and ReLU activation. Notably, feature maps are downsampled in the second, third, and fourth layers using a stride of 2 in the first convolutional layer of each set. Following this, adaptive average pooling is used in the global average pooling layer to reduce each channel's feature maps to a single value before flattening. Finally, the combined features pass through a fully connected layer that produces the output results. Modifications made to the original model architecture involve removing batch normalization layers to enable training under differential privacy constraints, as these layers are incompatible with differential privacy due to their sharing of information across different training examples within a single minibatch, thereby compromising privacy guarantees [44].

CNN-Transformer Network (CTN) Natarajan et al. [17] proposed a wide and deep Transformer-based neural network for classifying 12-lead ECG sequences into 27 distinct cardiac abnormality classes. Their model was among the top-performing entries in the CinC 2021 Challenge. To broaden the applicability of our study by comparing different deep learning architecture with differential privacy, we implemented the same architectural framework with a key modification: replacing the batch normalization layers with group normalization layers in order to be compatible with DP requirements. The final architecture consists of three main components: an embedding network, a Transformer encoder, and a classification head. The embedding network comprises six 1D convolutional layers, each followed by group normalization (using 32 groups) and ReLU activation. This network processes the raw ECG signal into a sequence of 256-dimensional embeddings, achieving an approximate downsampling factor of 20. These embeddings are then passed into a Transformer encoder with 8 layers, each containing 8 attention heads and a feed-forward network with 2048 units. Positional encodings are added to the input embeddings to retain temporal information. The Transformer output is globally average-pooled to produce a single 256-dimensional representation. This vector is passed through a fully connected layer that reduces the dimensionality to 64, followed by ReLU activation and dropout (rate 0.2), and finally through another fully connected layer to produce the output logits for the classification task.

3.4. Modeling with handcrafted ECG features

We followed the methodology of Natarajan et al. [17], who initially extracted more than 300 features from the lead II of ECG recordings. These included heart rate variability features (time, frequency domain and non-linear), as well as morphological features. Using a random forest model, they assessed feature importance and identified the top 20 ECG-derived features most predictive of cardiac abnormalities. Building on their work, we adopted these top features and supplemented them with two demographic variables (age and gender) resulting in a total of 22 handcrafted ECG features we used for training.

Logistic Regression (LR): As a baseline for research question , we implemented a multinomial logistic regression model. The model ingests the 22-dimensional feature vector and applies an affine transformation to compute separate linear scores for each class. A sigmoid

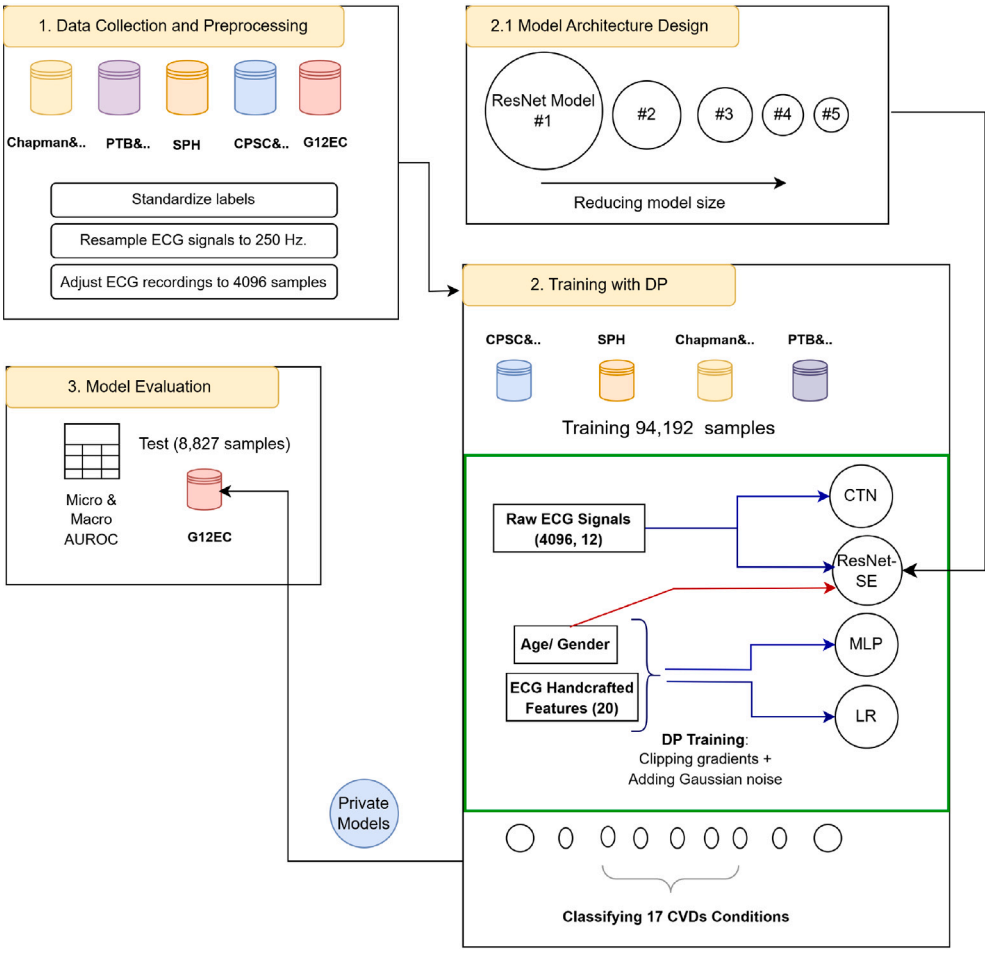

Fig. 1. End-to-end workflow of the ECG classification framework. The process includes (1) data collection and preprocessing across five public datasets, (2) Combining four data sources for training different model architectures training under DP using raw ECG signals, demographic information, and handcrafted features, and (3) model evaluation on a held-out test set (G12EC) using Micro and Macro-average AUC. The architecture exploration as in (2.1) includes a progressively smaller ResNet-SE models.

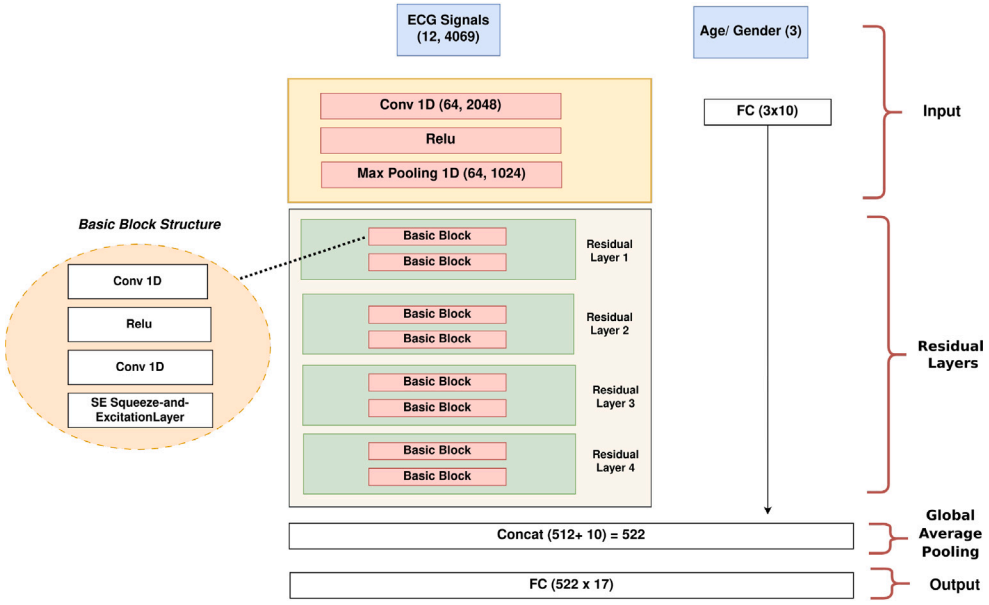

Fig. 2. ResNet-SE model architecture [15].

activation is then used at the output layer to enable independent binary decisions for each label. We optimized the model using binary cross-entropy loss and stochastic gradient descent.

Multilayer Perceptron (MLP): Our MLP baseline trained only on the handcrafted ECG features, rather than ECG raw signals. It consists of four fully connected layers: an input layer of size 22 followed by hidden layers of 512, 128 and 64 units. After each of the first three linear layers we apply a ReLU activation and a dropout for regularization. The final output layer employs a sigmoid activation to generate independent probability scores for each class, making the model suitable for multi-label classification of ECG-derived features.

3.5. Performance evaluation measures

To evaluate the model performance on our multi-label classification task, we used two metrics: macro-average and micro-average area under the receiver operating characteristics (AUC) [45]. Macro-average AUC is calculated by averaging the area under the curves (AUCs) for each label, ensuring that each label contributes equally to the overall metric. This approach gives a balanced assessment of the model, more so in healthcare applications, where missing even rare conditions can be critical. Micro-average AUC, on the other hand, combines all the predictions and actual labels across all the labels to compute a combined true positive rate (TPR) and false positive rate (FPR). This results in an AUC based on a unified ROC curve, demonstrating how well the model performs in the conditions that occur most frequently. In addition to these overall metrics, we calculated the AUC score for each of the individual classes to measure the model's performance in classifying both frequent and rare conditions.

4. Experiments

We conducted our experiments for model training using two NVIDIA Titan RTX GPUs. The development environment included PyTorch version 2.2.2 with CUDA toolkit 10.1 (cu121) on Python 3.10.14. For differential privacy implementations, Opacus version 1.4.0 was utilized to handle privacy parameters (ϵ, δ). Experiment tracking and visualization were managed through Weights & Biases (Wandb) version 0.16.6.

We conducted our experiments using a training set of 94,192 samples, which included data from SPH, PTB & PTBXL, Chapman-Shaoxing & Ningbo, and CPSC & CPSC-Extra. The G12EC dataset, consisting of 8,827 samples, was used as our test set. We chose this split to better reflect real-world scenarios, where models are typically trained on data from certain hospitals and then deployed in new, unseen environments. We selected G12EC as the test set because it provides good coverage of all the labels, ensuring a comprehensive evaluation of the model. Fig. 1 detailed the full steps of our experiment, starting from data preprocessing and prepareting and ending with model training and evaluation. To evaluate model stability and enable performance comparisons between different architectures trained under DP, we constructed 95% confidence intervals based on repeated experiments. Each model was trained ten times using the same training and test datasets, with a different random seed used each run to introduce variability. We then computed the mean and standard deviation of the resulting performance scores and applied the t-distribution to estimate the confidence intervals. These statistics were used to derive the lower and upper bounds, capturing the uncertainty in model performance, which is substantially amplified by the noise introduced by the DP mechanism.

4.1. Training with differential privacy

In this section, we address our first research question (). To evaluate the performance of models trained with and without differential privacy, we explored various model architectures, including the ResNet-SE model, as illustrated in Fig. 2, as well as a transformer-based model (CTN). These models were trained on raw ECG signals. Additionally, we trained two baseline models: MLP and logistic regression models, which were based on extracted handcrafted ECG features, along with age and gender data. For training the ResNet-SE model without differential privacy, we used the default training configurations, which yielded the best results during the CINC challenge 2021. Specifically, the classifier was trained with a batch size of 64 and a learning rate of 0.003. We ran the model for 20 epochs.

When training the ResNet-SE model under DP constraints, additional DP-specific parameters, such as δ, privacy budget, and clipping norm, must be incorporated into the training process. Tuning these DP-specific parameters, along with No-DP-related parameters such as batch size, learning rate, and number of iterations, is crucial because they jointly influence the utility and privacy of the trained model.

Three possible strategies for hyperparameter tuning under DP constraints have been suggested in the literature [16, Sec. 5.4.1]. These involve optimizing over three interdependent objectives: (1) model accuracy, (2) privacy cost (ϵ), and (3) computational cost. Each strategy prioritizes optimizing one of these objectives while treating the other two as constraints. For tuning under computational constraints, the proposed approach includes scaling the batch size and number of epochs to computational limits, tuning the clipping norm, and performing random searches over the learning rate for different clipping norms based on some fixed privacy budgets.

In our study, we adopted a similar approach by tuning the hyperparameters under computational constraints for different specified ϵ targets (1, 10, 100). This was achieved through a random search over predefined hyperparameter values for the ResNet-SE model, as detailed in Table 4.

Due to computational and memory constraints, conducting a comprehensive random search across all parameter combinations for all combined hospitals was infeasible. Instead, we performed preliminary hyperparameter tuning on a single dataset, *Chapman-Shaoxing and Ningbo*, which has the largest sample size of 43,560 patients. For this tuning, a small validation set was randomly selected from the training data. Hyperparameters were varied across a predefined range of values, and for each combination, the model was trained on the remaining training data and evaluated on the validation set. The objective was to minimize the validation loss, and the hyperparameter configuration that achieved the lowest validation loss was selected for subsequent training on the full dataset. We found that the optimal hyperparameter values were a batch size of 1000, a clipping norm of 1, and a learning rate of 0.003. Using these optimal hyperparameters, we trained the final models on the complete dataset, which includes SPH, PTB & PTBXL, Chapman-Shaoxing & Ningbo, and CPSC & CPSC-Extra, and tested them on the G12EC dataset. For the privacy parameter δ, we configured it as $\frac{1}{N}$, where $N = 94{,}192$, representing the total number of training samples. Additionally, we manually tuned the number of training epochs by evaluating the models across different values [5, 10, 20, 100]. The table in 4 presents the set of hyperparameters and the best choices for each parameter after tuning for each model architectures. However, for CTN and MLP we manually tuned the parameters, while for the logistic regression, which is a lightweight model with acceptable training time, we performed hyperparameter tuning via grid search on the Chapman–Ningbo dataset under a privacy budget of $\epsilon = 1$. We evaluated batch size ratios of 0.05, 0.1, 0.2, and 0.5 (corresponding to 5%, 10%, 20%, and 50% of the training set per batch), as well as various learning rates and epoch counts. However, hyperparameter tuning can lead to privacy leakage because it involves repeatedly accessing sensitive data to evaluate different hyperparameter settings, which cumulatively can

Table 4
Hyperparameter search values for all model architectures trained with and without DP.

Model	Hyperparameter	Search values	Best choice
ResNet-SE	Learning Rate	1×10^{-4}, 3×10^{-2}, 1×10^{-2}, 1×10^{-1}	3×10^{-2}
	Batch Size	200 to 2000 (step size = 200)	1000
	Gradient Clipping Norm	1×10^{-3}, 1×10^{-2}, 1×10^{-1}, 1, 10	1
CTN	Batch Size	32, 400, 500	400
	Epochs	5, 6, 8, 10, 20	5
	Gradient Clipping Norm	1×10^{-1} 1, 10	1
LR	Learning Rate	0.1, 1, 5, 10, 20, 50	20
	Batch Size	0.05, 0.1, 0.2, 0.5 (e.g 0.05 mean batch size = 5% of the training samples)	0.5
	Epochs	10, 20, 50, 100, 200, 500	500
MLP	Dropout	0, 0.5	0.5
	Batch Size	32, 64, 128	64
	Epochs	7, 10, 20, 30, 40	7

Table 5
Macro-average and Micro-average AUC scores with 95% confidence intervals for four different models (ResNet-SE, CTN, MLP, and Logistic Regression) evaluated under varying differential privacy budgets. Performance is reported for both non-private (No-DP) and private training settings.

Metric	Privacy budget	Models			
		ResNet-SE	CTN	MLP	Logistic Reg.
Macros	No-DP	0.904 (0.902–0.906)	0.880 (0.874–0.887)	0.713 (0.711–0.715)	0.719 (0.719–0.720)
	$\epsilon = 100$	0.844 (0.838–0.850)	0.667 (0.661–0.672)	0.695 (0.689–0.702)	0.719 (0.718–0.720)
	$\epsilon = 10$	0.786 (0.777–0.796)	0.644 (0.639–0.650)	0.702 (0.697–0.708)	0.719 (0.719–0.720)
	$\epsilon = 1$	0.754 (0.743–0.766)	0.597 (0.574–0.620)	0.704 (0.702–0.706)	0.708 (0.705–0.711)
Micro	No-DP	0.923 (0.920–0.927)	0.907 (0.901–0.913)	0.834 (0.830–0.837)	0.810 (0.808–0.811)
	$\epsilon = 100$	0.863 (0.855–0.871)	0.743 (0.738–0.747)	0.821 (0.815–0.828)	0.812 (0.811–0.814)
	$\epsilon = 10$	0.815 (0.798–0.832)	0.727 (0.724–0.731)	0.817 (0.811–0.824)	0.812 (0.810–0.814)
	$\epsilon = 1$	0.796 (0.780–0.812)	0.699 (0.688–0.711)	0.823 (0.820–0.825)	0.773 (0.760–0.787)

expose private information. In our work, we did not account for privacy loss that can occur through hyperparameter selection, as addressing the model selection problem in DP was outside the scope of our work. Works such as Liu et al. [46] and Shubhankar et al. [47] have proposed possible solutions for addressing this issue.

The results presented in Table 5 indicate that all models after hyperparameter tuning achieve their highest macro- and micro-average AUC scores when no differential privacy is applied (ResNet-SE: 0.904/0.923; CTN: 0.880/0.907; MLP: 0.713/0.834; Logistic Regression: 0.719/0.810). As the privacy budget decreases (indicating stronger privacy), the performance of each model declines, although the extent of this decline varies significantly among the models. The ResNet-SE model shows a moderate decrease in performance (macro: −0.06 at $\epsilon = 100$; −0.15 at $\epsilon = 1$; micro: −0.06 at $\epsilon = 100$; −0.13 at $\epsilon = 1$). In contrast, the CTN architecture, which has 13,622,149 parameters, experiences the most considerable drops (macro: −0.20 at $\epsilon = 100$; −0.28 at $\epsilon = 1$; micro: −0.16 at $\epsilon = 100$; −0.21 at $\epsilon = 1$). This suggests that a model's capacity can amplify the negative effects of gradient clipping and noise. Specifically, the CTN model is 1.54 times larger than the ResNet-SE, and the total amount of noise introduced is proportional to the number of model parameters. This may explain why the CTN model was most affected when trained under DP. Conversely, the baseline models trained on handcrafted ECG features (MLP and Logistic Regression) only exhibit slight degradations (macro drops ≤ 0.02 at $\epsilon = 1$; micro drops ≤ 0.04 at $\epsilon = 1$), but their performance is low to start with. These trade-offs demonstrate the challenge of maintaining model performance under strict privacy constraints even after tuning the hyperparamters, and highlight the need to explore alternative approaches, such as redesigning the model architecture to accommodate better differential privacy, which is discussed in Section 4.2. Overall, the ResNet-SE architecture consistently outperforms others across various privacy budgets.

4.2. Model size impact on DP

To address our second research question(), this section explores how variations in model size influence the effectiveness of models trained with DP. We chose the ResNet-SE architecture for this analysis because it outperformed all other models in the non-private setting. Specifically, we aimed to investigate the effects of reducing model complexity to achieve a balance between privacy and utility. In particular, we focused on adjusting the number of output channels (filters) in the residual layers of the original ResNet-SE model while maintaining the same structure in the other layers. Fig. 3 highlights the architectural differences across different simplified versions of ResNet-SE models. In the baseline architecture refereed as ResNet-SE #1, the number of output channels in RLs 1 through 4 is 64, 128, 256, and 512, respectively. This setup was designed to extract complex features from 12-lead ECG signals. ResNet-SE #2 halving the number of output channels in these layers to 32, 64, 128, and 256, respectively. Similar reductions are applied in ResNet-SE #3 and #4. The most simplified architecture, ResNet-SE #5, uses only two output channels for each residual layer, representing the minimal configuration. Reducing the number of output channels directly affects the model's complexity and, crucially, the amount of noise added in DP training. By decreasing the number of output channels, the model has fewer learnable parameters, resulting in smaller gradient vectors. Since the total noise added is proportional to the number of parameters (e.g, the dimensionality of the gradient), a model with fewer parameters requires less cumulative noise to be injected during training [16].

Fig. 4 presents the macro and micro-average AUC values for five different ResNet-SE sizes, labeled #1 through #5, in addition to CTN model across various privacy settings defined by ϵ values of 1, 10, 100, and No-DP. The results indicate that both macro-average and micro-average AUC scores decrease as privacy settings become stricter. This trend is particularly evident in the red bars, representing the lowest privacy budget $\epsilon = 1$, where the impact of added noise on model accuracy is most pronounced. In contrast, the blue bars, which represent the models trained without DP, generally show the highest scores,

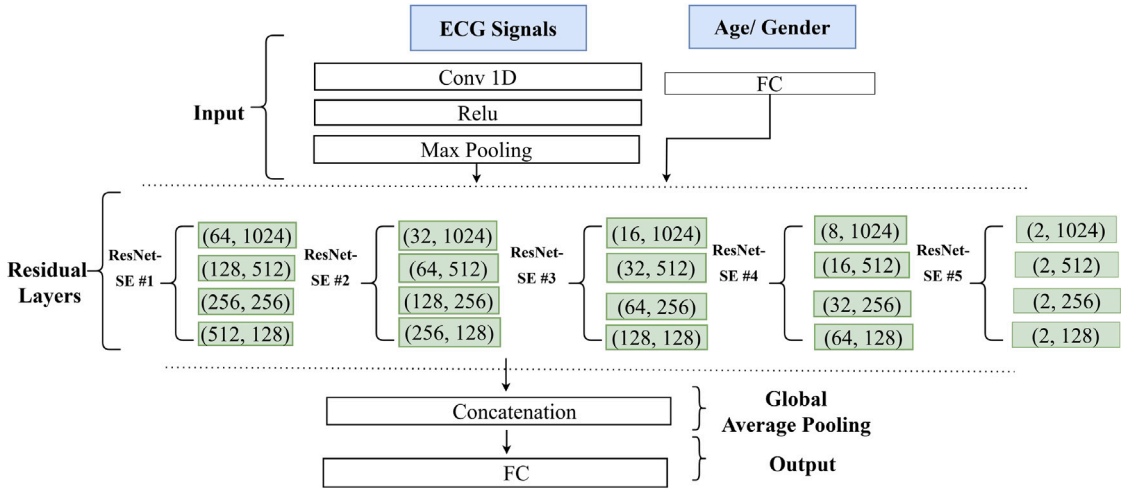

Fig. 3. Architectures for the five models evaluated, showing variations in the number of output channels in the residual layers (RLs). ResNet-SE #1 represents the baseline architecture.

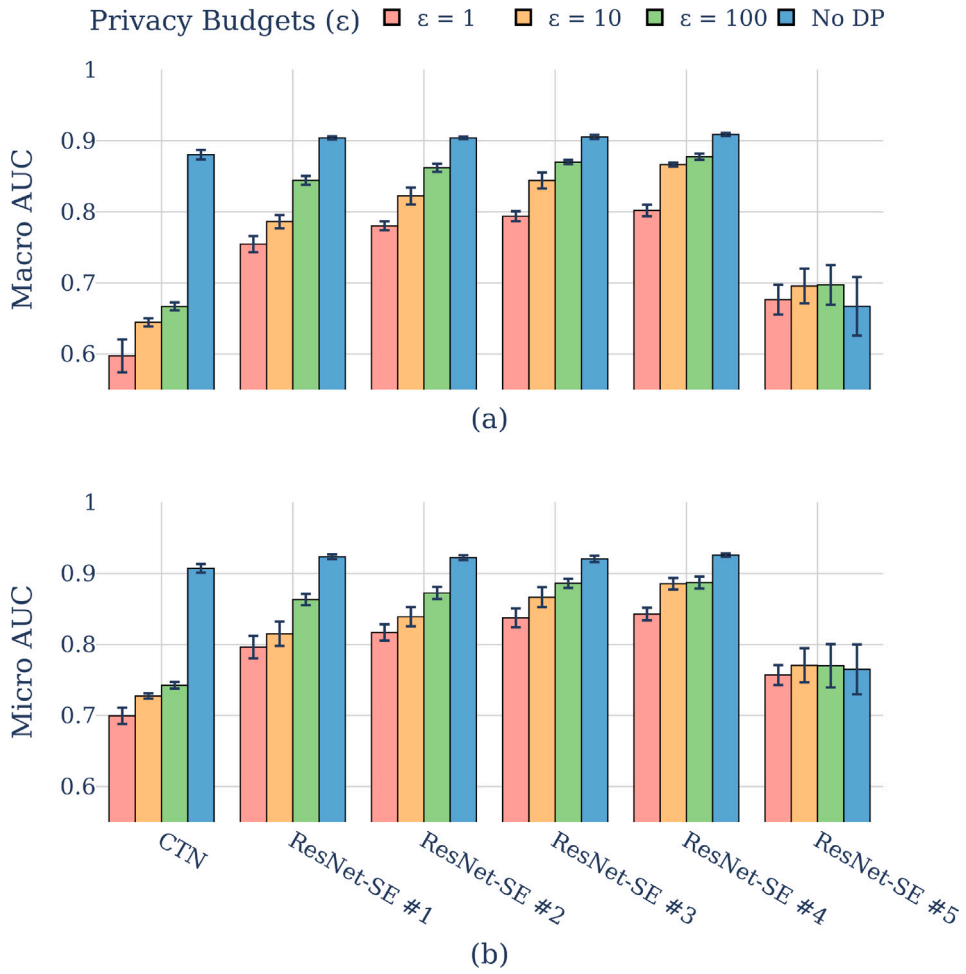

Fig. 4. Comparison of macro-average and micro-average AUC scores with 95% confidence intervals for all ResNet-SE models and CTN across different privacy budgets.

pointing to the trade-off between privacy and model performance. ResNet-SE #5 initially performs poorly when trained without DP, likely due to its simplified architecture. However, when DP is applied, its performance improves slightly. This suggests that the combination of architectural simplicity and DP-induced noise can enhance the model's generalizability. As noted by [48], DP can help reduce overfitting by injecting carefully calibrated noise in certain cases.

The error bars for ResNet-SE #5, representing the 95% confidence interval, exhibit significantly higher variability compared to other architectures across all privacy levels.

All other model architectures demonstrated relatively stable performance when trained without DP. However, once DP was applied, performance variability increased across the board, becoming especially pronounced at $\epsilon = 1$ and $\epsilon = 10$. This is due to the greater amount of noise introduced under stricter privacy budgets, which leads to less stable training and highlights the inherently stochastic nature of differentially private learning.

Among all models, ResNet-SE #4 consistently achieves the best performance in both macro-average and micro-average AUC scores. It also demonstrates strong robustness to privacy constraints, with only a 0.04 drop in both metrics at $\epsilon = 10$ compared to training without DP. In contrast, the CTN model performs competitively without DP (AUC = 0.88), coming close to ResNet-SE #4. However, under stronger privacy settings, its performance declines significantly. For example, all macro-average values fall below 0.67 across the evaluated privacy budgets. This drop is likely due to the model's large size and complexity which effected by the DP noise. While ResNet-SE #4 has only 142,934 trainable parameters, the CTN has 13,622,149 approximately 95 times more. As a result, the noise added to the CTN during each training epoch is much larger, which harms performance. This highlights the value of simplifying models to achieve better privacy-utility trade-offs.

To examine the effects of model complexity and DP on macro- and micro-average AUC, linear regression models were fitted using the 'statsmodels' Python package. Model complexity was encoded as a binary variable, with ResNet-SE #1 representing high complexity and ResNet-SE #4 representing low complexity. Another binary variable was used to encode whether DP was applied or not. We modeled the AUC as the response variable, and model complexity, DP and their interaction term as explanatory variables. Statistical significance of the regression coefficients was assessed using Student's t-tests with Bonferroni correction applied for multiple comparisons and significance level $\alpha = 0.05$. The results were the same for both macro- and micro-average AUC. While alone the effect of complexity was not statistically significant, the presence of DP was associated with a significant decrease in AUC. Most notably, a significant interaction between complexity and DP was observed, indicating that the reduction in AUC due to DP was more pronounced in the high-complexity model.

4.3. DP impact on the underrepresented classes

To address our third research question (), we evaluated the AUC performance of ResNet-SE #4, which demonstrated the best results under privacy constraints, as illustrated in Fig. 4. We compared the AUC scores for each class among models trained with and without differential privacy under various privacy budgets. To account for variability due to training randomness, we repeated all experiments ten times with different random seeds for each privacy budget ($\epsilon = 1$, $\epsilon = 10$, $\epsilon = 100$, and No-DP). The results, presented in Table 6, include the mean and standard deviation (AUC $mean \pm std$) for each label under the three privacy settings. We ordered all classes in descending order based on class size.

our analysis show that the No-DP model consistently outperformed the DP models across all individual classes, as highlighted in red for the mean AUC scores. Interestingly, for the well-represented class "Sinus Rhythm," the AUC scores of the DP models closely aligned with those of the No-DP model. In particular, the DP model with $\epsilon = 100$ even outperformed the No-DP model for this class, achieving a mean AUC of 0.92, which is a 0.02 increase over the No-DP model's score of 0.90. Additionally, for the class "T Wave Abnormal", there was also an increase of 0.02 when $\epsilon = 10$.

Overall, the performance degradation in AUC was more pronounced for the DP model with $\epsilon = 1$, particularly in classes with smaller class sizes. This is due to the increased noise required for stricter privacy. Notably, AUC declines were not strictly correlated with class size. For example, the class "Left Bundle Branch Block", despite being the most underrepresented in the training data with a class size of 1.19%, achieved a mean AUC of approximately 0.86 for the DP model with $\epsilon = 1$ and 0.96 for the No-DP model, representing a 0.10 drop. In contrast, classes such as "Low QRS Voltage" and "Premature Atrial Contraction", with higher class sizes of 1.64% and 3.10%, respectively, experienced substantial performance drops in the DP models. Specifically, for "Low QRS Voltage", the DP model with $\epsilon = 1$ achieved a mean AUC of 0.57, compared to 0.76 in the No-DP model, representing a performance drop of 0.19. Similarly, for "Premature Atrial Contraction", the DP model with $\epsilon = 1$ achieved a mean AUC of 0.63, compared to 0.91 in the No-DP model, resulting in a performance drop of 0.28.

Lastly, we observed that the variation in AUC performance was more pronounced for models trained with $\epsilon = 1$ compared to those with higher privacy budgets. This is evident in the standard deviations highlighted in blue, where all classes except the top two most represented ones have higher standard deviations compared to models trained with higher privacy budgets. This increased variability can be attributed to the high level of added noise and the inherent randomness in the DP training process, leading to a wide range of possible outcomes in model performance across different training runs. Overall, strong DP had minimal impact on the four largest classes, with an AUC reduction of approximately ≤ 0.01. However, smaller classes particularly those with a class ratio below 9% experienced more substantial declines in performance. To investigate whether the size of the class is correlated with the performance degradation caused by DP, we analyze the drop in AUC between the non-private and private models (with $\epsilon = 1$) for each class. Using the Kendall rank correlation coefficient, we calculated the relationship between class size and drop. Our analysis resulted in Kendall's τ of -0.508, indicating a moderate to strong negative correlation. This suggests that larger classes tend to experience less performance degradation under DP, while smaller classes are more adversely affected. This finding highlights an important consideration for fairness when implementing DP: minority or underrepresented classes may suffer disproportionately under privacy constraints.

Fig. 5 evaluates the influence of model architectures on performance across individual classes, comparing ResNet-SE #1 with ResNet-SE #4. When trained without differential privacy, as depicted in Fig. 5(a), both models exhibited nearly identical performance across all classes. This parity indicates a baseline comparability in their learning capabilities under non-restrictive conditions. However, under strict privacy constraint ($\epsilon = 1$), as shown in Fig. 5(b), notable differences emerged. ResNet-SE #4 outperformed ResNet-SE #1 in 11 classes, demonstrating better AUC scores. In the remaining 6 classes, ResNet-SE #1 showed marginal improvements. This differential suggests that the architectural simplifications implemented in ResNet-SE #4 enhanced its ability to maintain performance integrity under the noise introduced by strict DP measures on the individual classes.

5. Discussion and conclusion

In this paper, we studied the trade-off between model accuracy and privacy for both state-of-the-art ResNet-SE and CTN models. They showed a significant drop in micro and macro-average AUC performance under strict privacy settings. However, through model size reduction from ResNet-SE #1 to ResNet-SE #4, we were able to improve this trade-off. For example, with a privacy budget of $\epsilon = 1$, the macro-average AUC improved from 0.75 to 0.80, representing a 0.05

Table 6
Mean ± Std (AUC Scores) for ECG Conditions under different privacy budgets for the ResNet-SE #v4 with 10 multiple runs with different random seeds on the same test set. Blue numbers indicate the highest standard deviation, while red numbers highlight the highest mean scores.

Condition	$\epsilon = 1$	$\epsilon = 10$	$\epsilon = 100$	No DP
Sinus Rhythm	0.90 ± 0.0105	0.91 ± 0.0127	**0.92** ± 0.0135	0.90 ± 0.0139
Sinus Bradycardia	0.98 ± 0.0053	0.98 ± 0.0157	0.96 ± 0.0331	**0.99** ± 0.0019
T Wave Abnormal	0.78 ± 0.0145	**0.80** ± 0.0114	0.77 ± 0.0213	0.78 ± 0.0347
Sinus Tachycardia	0.98 ± 0.0059	0.99 ± 0.0022	0.99 ± 0.0027	**0.99** ± 0.0009
Atrial Flutter	0.84 ± 0.0245	0.88 ± 0.0076	0.89 ± 0.0137	**0.91** ± 0.0068
Left Axis Deviation	0.89 ± 0.0179	0.92 ± 0.0068	0.92 ± 0.0088	**0.94** ± 0.0074
Atrial Fibrillation	0.84 ± 0.0303	0.89 ± 0.0108	0.90 ± 0.0121	**0.93** ± 0.0021
Sinus Arrhythmia	0.61 ± 0.0530	0.77 ± 0.0432	0.81 ± 0.0408	**0.95** ± 0.0159
T Wave Inversion	0.63 ± 0.0083	0.64 ± 0.0116	0.65 ± 0.0053	**0.68** ± 0.0020
1st Degree AV Block	0.79 ± 0.0183	0.91 ± 0.0148	0.94 ± 0.0082	**0.97** ± 0.0035
Right Bundle Branch Block	0.90 ± 0.0171	0.94 ± 0.0103	0.94 ± 0.0087	**0.95** ± 0.0078
Premature Atrial Contraction	0.63 ± 0.0172	0.79 ± 0.0172	0.85 ± 0.0148	**0.91** ± 0.0063
Incomplete Right Bundle Branch Block	0.72 ± 0.0332	0.87 ± 0.0097	0.89 ± 0.0105	**0.91** ± 0.0094
Left Anterior Fascicular Block	0.87 ± 0.0280	0.91 ± 0.0108	0.93 ± 0.0034	**0.95** ± 0.0057
Low QRS Voltage	0.57 ± 0.0370	0.65 ± 0.0217	0.67 ± 0.0142	**0.76** ± 0.0190
Right Axis Deviation	0.83 ± 0.0703	0.93 ± 0.0148	0.93 ± 0.0117	**0.96** ± 0.0022
Left Bundle Branch Block	0.86 ± 0.0426	0.94 ± 0.0114	0.95 ± 0.0078	**0.96** ± 0.0026

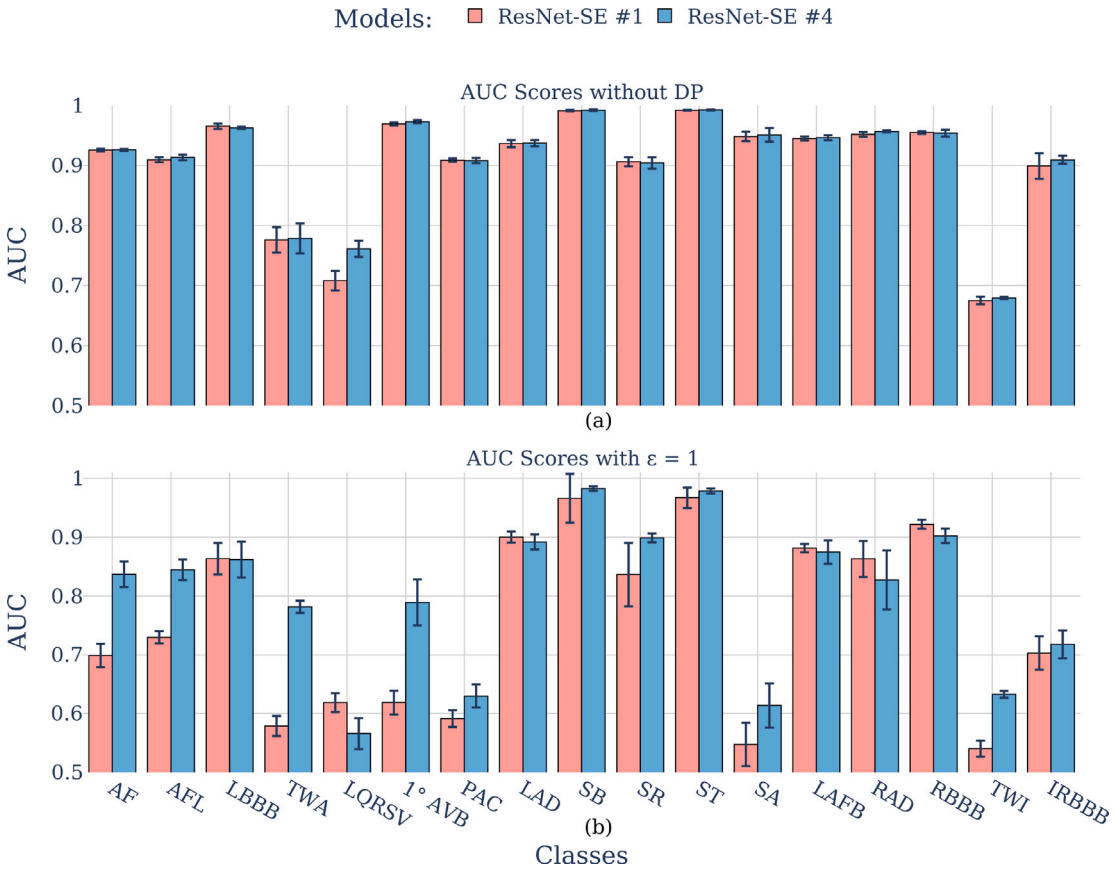

Fig. 5. Comparison of AUC scores for individual classes between ResNet-SE #1 and ResNet-SE #4 under different privacy budgets: (a) without differential privacy and (b) with strict privacy budget ($\epsilon = 1$). Error bars indicate 95% confidence intervals.

improvement. Similarly, with a privacy budget of $\epsilon = 10$ and $\epsilon = 100$, the macro-average AUC was improved by 0.08 and 0.04 respectively. Therefore, considering a modest privacy budget $\epsilon = 10$, we could have a private model with favorable performance with only 0.03 drop in macro- and micro-average performance compared to the non-private model, in which it can finally be deployed while minimizing the risk of leaking patient privacy information.

In general, it was observed that the size of a model significantly impacts the utility of DP models. Transformer models, in particular, tend to have a large number of parameters and high-dimensional gradients, which can result in substantial noise accumulation when trained under DP constraints. This accumulation can destabilize the training process and affect attention scores. In our case, the CTN model, which has more than 13.6 million parameters, exhibited a severe performance drop. The macro-average AUC decreased from 0.88 to 0.67, and the micro-average AUC decreased from 0.91 to 0.74 at $\epsilon = 100$. Performance dropped even further at $\epsilon = 10$ and $\epsilon = 1$. However, recent work by [49] systematically addresses the unique challenges associated with training Transformer models under differential privacy. The authors identify two critical issues: the attention distraction phenomenon, where noise disproportionately affects rare tokens and skews attention scores, and the incompatibility of standard Transformer architectures with efficient

per-sample gradient clipping. To address these problems, they propose the Re-Attention Mechanism, which corrects attention bias by tracking and adjusting for activation variance, and Phantom Clipping, which facilitates efficient gradient clipping while maintaining shared embeddings. Their modular approach effectively bridges the gap between DP-Transformers and traditional differential privacy training methods, enhancing both training stability and utility.

Other researchers have also argued that overparameterized models may perform poorly with DP-SGD and recommended reducing the dimensionality of updates, either explicitly or implicitly. This can be achieved by using smaller models, applying dimensionality reduction techniques, or utilizing handcrafted features [25,50,51]. Conversely, [52] demonstrated that large, over-parameterized models can still achieve strong performance when proper hyperparameter tuning and architectural adjustments are applied. By reducing regularization and making other changes, they achieved state-of-the-art results on CIFAR-10, with a performance boost of approximately 10%. This suggests that exploring both small and large models with careful tuning is important to determine the best architecture for DP settings. [53] have suggested that instead of reusing architectures optimized for No-DP settings, models may need to be redesigned specifically for DP training using automated neural architecture search.

This study also outlined that strict private models exhibit more variability and can be difficult to replicate across different random seeds. Berrada et al. [54] investigated per-class disparities by training 50 different Wide Residual Network (WRN) models under differential privacy at each ϵ on CIFAR-10, using different random seeds. They found that the variation was much higher in private models compared to non-private ones. The variation in class-conditional accuracy was due to private models over- and under-predicting certain classes. However, using early stopping or Exponential Moving Average (EMA) checkpoints produced more robust disparity measurements and reduced variations in private models without incurring any additional privacy cost.

6. Limitations and future works

In our study, we did not account for the privacy cost associated with hyperparameter tuning, but in deployment scenarios, it must be considered. We conducted approximately 270 experiments to optimize the hyperparameters of ResNet-SE #1 alone, with even more runs dedicated to tuning the hyperparameters of its simplified model variants, as well as other model architectures such as CTN, MLP, and logistic regression. All these runs increase the privacy budget and must be accounted for when calculating the final privacy cost. The simplest method for accounting for privacy costs during hyperparameter tuning is through sequential composition, which involves calculating the cumulative privacy cost by summing the individual (ϵ) and (δ) values from each tuning run. However, there are improvements over sequential composition, such as using the Exponential mechanism [11] or Randomized Number of Trials [55]. Ideally, hyperparameter tuning should be performed on public datasets [16], and the optimal parameters can then be applied to train the model under DP constraints to avoid any privacy leakage.

In this study, we limited our selection of models to those compatible with DP-SGD, focusing on deep learning models, logistic regression, and MLPs that provide consistent and comparable privacy guarantees when implemented using the PyTorch-based Opacus framework. However, future research could explore more advanced non-deep-learning models that support differential privacy, such as Differentially Private Support Vector Machines (DP-SVM) [56] or Differentially Private Gradient Boosted Decision Trees (DP-GBDT) [57], and train them on our handcrafted ECG features. To make the results more generalizable, future work could conduct tuning on multiple datasets that represent a wider range of data sources. Cross-validation across different hospitals would also provide more insight into how well the model generalizes to new, unseen environments. Another promising direction for future research would be to explore federated learning in combination with DP. Federated learning allows for model training across multiple hospitals without the need to centralize patient data, thereby increasing data security.

CRediT authorship contribution statement

Zoher Orabe: Writing – original draft, Visualization, Software, Methodology, Investigation. **Antti Vasankari:** Writing – original draft, Software, Investigation. **Tapio Pahikkala:** Writing – review & editing, Supervision, Methodology, Conceptualization. **Matti Kaisti:** Writing – review & editing, Supervision, Methodology, Conceptualization. **Antti Airola:** Writing – review & editing, Supervision, Methodology, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work has been supported by Research Council of Finland (grant 358868).

Data availability

The complete codebase used to build the training pipeline including data preparation, model training, and evaluation of ECG models with DP is publicly available on GitHup. The repository includes all necessary scripts, configuration files, and detailed instructions to reproduce the experiments presented in this paper, in addition to the links for downloading the ECG datasets. The code can be accessed at: https://github.com/UTU-Health-Research/dl-ecg-dp-classifier/.

References

[1] World Health Organization: WHO, Cardiovascular diseases, 2019, URL https://www.who.int/health-topics/cardiovascular-diseases.

[2] Y. Li, J. hao Luo, Q. yun Dai, J.K. Eshraghian, B.W.-K. Ling, C. yan Zheng, X. li Wang, A deep learning approach to cardiovascular disease classification using empirical mode decomposition for ECG feature extraction, Biomed. Signal Process. Control. 79 (2023) 104188, http://dx.doi.org/10.1016/j.bspc.2022.104188.

[3] P. Rajpurkar, A.Y. Hannun, M. Haghpanahi, C. Bourn, A.Y. Ng, Cardiologist-level arrhythmia detection with convolutional neural networks, 2017, CoRR abs/1707.01836.

[4] M. Fredrikson, S. Jha, T. Ristenpart, Model inversion attacks that exploit confidence information and basic countermeasures, in: Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security, CCS '15, Association for Computing Machinery, New York, NY, USA, 2015, pp. 1322–1333, http://dx.doi.org/10.1145/2810103.2813677.

[5] P. Rockenschaub, et al., The impact of multi-institution datasets on the generalizability of ICU risk prediction models, Crit. Care Med. (2024).

[6] C. Fang, A. Dziedzic, L. Zhang, L. Oliva, A. Verma, F. Razak, N. Papernot, B. Wang, Decentralised, collaborative, and privacy-preserving machine learning for multi-hospital data, EBioMedicine 101 (2024) 105006, http://dx.doi.org/10.1016/j.ebiom.2024.105006.

[7] E.A.P. Alday, A. Gu, A.J. Shah, C. Robichaux, A.-K.I. Wong, C. Liu, F. Liu, A.B. Rad, A. Elola, S. Seyedi, Q. Li, A. Sharma, G.D. Clifford, M.A. Reyna, Classification of 12-lead ECGs: the PhysioNet/Computing in cardiology challenge 2020, Physiol. Meas. 41 (12) (2020) 124003, http://dx.doi.org/10.1088/1361-6579/abc960.

[8] T. Leinonen, D. Wong, A. Vasankari, A. Wahab, R. Nadarajah, M. Kaisti, A. Airola, Empirical investigation of multi-source cross-validation in clinical ECG classification, Comput. Biol. Med. 183 (2024) 109271, http://dx.doi.org/10.1016/j.compbiomed.2024.109271.

[9] C. Dwork, F. McSherry, K. Nissim, A. Smith, Calibrating noise to sensitivity in private data analysis, in: Theory of Cryptography Conference, Springer, 2006, pp. 265–284.

[10] R. Cummings, D. Desai, The role of differential privacy in GDPR compliance, in: FAT'18: Proceedings of the Conference on Fairness, Accountability, and Transparency, vol. 20, 2018.
[11] M. Abadi, A. Chu, I. Goodfellow, H.B. McMahan, I. Mironov, K. Talwar, L. Zhang, Deep learning with differential privacy, in: Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, CCS '16, ACM, 2016, pp. 308–318, http://dx.doi.org/10.1145/2976749.2978318.
[12] E. Bagdasaryan, O. Poursaeed, V. Shmatikov, Differential privacy has disparate impact on model accuracy, in: H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. Fox, R. Garnett (Eds.), in: Advances in Neural Information Processing Systems, vol. 32, Curran Associates, Inc., 2019, pp. 15453–15462.
[13] V. Agrawal, S.V. Kalmady, V.M. Manoj, M.V. Manthena, W. Sun, M.S. Islam, A. Hindle, P. Kaul, R. Greiner, Federated learning and differential privacy techniques on multi-hospital population-scale electrocardiogram data, in: Proceedings of the 2024 8th International Conference on Medical and Health Informatics, ICMHI '24, Association for Computing Machinery, New York, NY, USA, 2024, pp. 143–152, http://dx.doi.org/10.1145/3673971.3673990.
[14] L. Zhang, J. Xu, A. Sivaraman, J. Deborah, P.K. Sharma, V. Pandi, A two-stage differential privacy scheme for federated learning based on edge intelligence, IEEE J. Biomed. Heal. Informatics (2023) 1–12, http://dx.doi.org/10.1109/jbhi.2023.3306425.
[15] Z. Zhao, H. Fang, S.D. Relton, R. Yan, Y. Liu, Z. Li, J. Qin, D.C. Wong, Adaptive lead weighted ResNet trained with different duration signals for classifying 12-lead ECGs, in: 2020 Computing in Cardiology, 2020, pp. 1–4, http://dx.doi.org/10.22489/CinC.2020.112.
[16] N. Ponomareva, H. Hazimeh, A. Kurakin, Z. Xu, C. Denison, H.B. McMahan, S. Vassilvitskii, S. Chien, A.G. Thakurta, How to DP-fy ML: A practical guide to machine learning with differential privacy, J. Artificial Intelligence Res. 77 (2023) 1113–1201, http://dx.doi.org/10.1613/jair.1.14649.
[17] A. Natarajan, Y. Chang, S. Mariani, A. Rahman, G. Boverman, S. Vij, J. Rubin, A wide and deep transformer neural network for 12-lead ECG classification, in: 2020 Computing in Cardiology, IEEE, 2020, pp. 1–4.
[18] R. Anil, B. Ghazi, V. Gupta, R. Kumar, P. Manurangsi, Large-scale differentially private BERT, in: Y. Goldberg, Z. Kozareva, Y. Zhang (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2022, Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 2022, pp. 6481–6491, http://dx.doi.org/10.18653/v1/2022.findings-emnlp.484.
[19] X. Li, F. Tramer, P. Liang, T. Hashimoto, Large language models can be strong differentially private learners, in: International Conference on Learning Representations, 2022.
[20] D. Kingma, J. Ba, Adam: A method for stochastic optimization, in: International Conference on Learning Representations, 2015.
[21] E.W. Lee, L. Xiong, V.S. Hertzberg, R.L. Simpson, J.C. Ho, Privacy-preserving sequential pattern mining in distributed EHRs for predicting cardiovascular disease, in: AMIA Summits on Translational Science Proceedings 2021, American Medical Informatics Association, 2021, p. 384.
[22] J. Chen, W.H. Wang, X. Shi, Differential privacy protection against membership inference attack on machine learning for genomic data, in: BIOCOMPUTING 2021: Proceedings of the Pacific Symposium, World Scientific, 2020, pp. 26–37.
[23] L. Fan, L. Bonomi, Mitigating membership inference in deep survival analyses with differential privacy, in: 2023 IEEE 11th International Conference on Healthcare Informatics, ICHI, IEEE, 2023, pp. 81–90.
[24] A. Ziller, D. Usynin, R. Braren, M. Makowski, D. Rueckert, G. Kaissis, Medical imaging deep learning with differential privacy, Sci. Rep. 11 (2021) http://dx.doi.org/10.1038/s41598-021-93030-0.
[25] C. Yan, H. Yan, W. Liang, M. Yin, H. Luo, J. Luo, DP-SSLoRA: a privacy-preserving medical classification model combining differential privacy with self-supervised low-rank adaptation, Comput. Biol. Med. 179 (2024) 108792.
[26] K. Gil, A. Vejar, Privacy-preserving framework for automated detection of arrhythmia in ECG data, J. Telecommun. Inf. Technol. (2025) http://dx.doi.org/10.26636/jtit.2025.fitce2024.2042.
[27] T.N. Islam, H. Imtiaz, A robust neural network for privacy-preserving heart rate estimation in remote healthcare systems, Heal. Anal. 5 (2024) 100329, http://dx.doi.org/10.1016/j.health.2024.100329.
[28] K. Weimann, T.O.F. Conrad, Federated learning with deep neural networks: A privacy-preserving approach to enhanced ECG classification, IEEE J. Biomed. Heal. Informatics 28 (11) (2024) 6931–6943, http://dx.doi.org/10.1109/jbhi.2024.3427787.
[29] E. Brophy, Synthesis of dependent multichannel ECG using generative adversarial networks, in: Proceedings of the 29th ACM International Conference on Information & Knowledge Management, 2020, pp. 3229–3232.
[30] L. Bonomi, Z. Wu, L. Fan, Sharing personal ECG time-series data privately, J. Am. Med. Informatics Assoc. 29 (7) (2022) 1152–1160.
[31] C. Dwork, A. Roth, The algorithmic foundations of differential privacy, Found. Trends Theor. Comput. Sci. 9 (3–4) (2014) 211–407, http://dx.doi.org/10.1561/0400000042.
[32] C. Dwork, A firm foundation for private data analysis, Commun. ACM 54 (1) (2011) 86–95, http://dx.doi.org/10.1145/1866739.1866758.
[33] I. Mironov, Rényi differential privacy, in: 2017 IEEE 30th Computer Security Foundations Symposium, CSF, 2017, pp. 263–275, http://dx.doi.org/10.1109/CSF.2017.11.
[34] J.P. Near, C. Abuah, Programming Differential Privacy, 2025, pp. 1–113, URL https://programming-dp.com/book.pdf.
[35] M.A. Reyna, N. Sadr, E.A.P. Alday, A. Gu, A.J. Shah, C. Robichaux, A.B. Rad, A. Elola, S. Seyedi, S. Ansari, H. Ghanbari, Q. Li, A. Sharma, G.D. Clifford, Will two do? Varying dimensions in electrocardiography: The PhysioNet/Computing in cardiology challenge 2021, in: 2021 Computing in Cardiology (CinC), vol. 48, 2021, pp. 1–4, http://dx.doi.org/10.23919/CinC53138.2021.9662687.
[36] M.A. Reyna, N. Sadr, E.A.P. Alday, A. Gu, A.J. Shah, C. Robichaux, A.B. Rad, A. Elola, S. Seyedi, S. Ansari, H. Ghanbari, Q. Li, A. Sharma, G.D. Clifford, Issues in the automated classification of multilead ecgs using heterogeneous labels and populations, Physiol. Meas. 43 (8) (2022) 084001, http://dx.doi.org/10.1088/1361-6579/ac79fd.
[37] A.L. Goldberger, L.A.N. Amaral, L. Glass, J.M. Hausdorff, P.C. Ivanov, R.G. Mark, J.E. Mietus, G.B. Moody, C.-K. Peng, H.E. Stanley, PhysioBank, PhysioToolkit, and PhysioNet, Circulation 101 (23) (2000) e215–e220, http://dx.doi.org/10.1161/01.CIR.101.23.e215.
[38] A large-scale multi-label 12-lead electrocardiogram database with standardized diagnostic statements, 2022, http://dx.doi.org/10.6084/m9.figshare.c.5779802.v1.
[39] E.Y.K. Ng, F. Liu, C. Liu, L. Zhao, X. Zhang, X. Wu, X. Xu, Y. Liu, C. Ma, S. Wei, Z. He, J. Li, An open access database for evaluating the algorithms of electrocardiogram rhythm and morphology abnormality detection, J. Med. Imaging Heal. Informatics (2018).
[40] R. Bousseljot, D. Kreiseler, A. Schnabel, Nutzung der EKG-signaldatenbank CARDIODAT der PTB über das internet, Biomed. Eng. / Biomed. Tech. 40 (s1) (1995) 317–318, http://dx.doi.org/10.1515/bmte.1995.40.s1.317.
[41] R.-D. Bousseljot, W. Samek, P. Wagner, T. Schaeffter, N. Strodthoff, PTB-XL, a large publicly available electrocardiography dataset (version 1.0.1) PhysioNet, 2020, http://dx.doi.org/10.13026/x4td-x982.
[42] J. Zheng, J. Zhang, S. Danioko, H. Yao, H. Guo, C. Rakovski, A 12-lead electrocardiogram database for arrhythmia research covering more than 10,000 patients, Sci. Data 7 (2020).
[43] J. Zheng, H. Chu, D. Struppa, J. Zhang, M. Yacoub, H. El-Askary, A. Chang, L. Ehwerhemuepha, I. Abudayyeh, A. Barrett, G. Fu, H. Yao, D. Li, H. Guo, C. Rakovski, Optimal multi-stage arrhythmia classification approach, Sci. Rep. 10 (2020) http://dx.doi.org/10.1038/s41598-020-59821-7.
[44] A.S. Shamsabadi, N. Papernot, Losing less: A loss for differentially private deep learning, 2022.
[45] J.A. Hanley, B.J. McNeil, The meaning and use of the area under a receiver operating characteristic (ROC) curve, Radiology 143 (1) (1982) 29–36, http://dx.doi.org/10.1148/radiology.143.1.7063747.
[46] J. Liu, K. Talwar, Private selection from private candidates, in: Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing, in: STOC 2019, Association for Computing Machinery, New York, NY, USA, 2019, pp. 298–309, http://dx.doi.org/10.1145/3313276.3316377.
[47] S. Mohapatra, S. Sasy, X. He, G. Kamath, O. Thakkar, The role of adaptive optimizers for honest private hyperparameter selection, 2021.
[48] C. Dwork, V. Feldman, M. Hardt, T. Pitassi, O. Reingold, A.L. Roth, Preserving statistical validity in adaptive data analysis, in: Proceedings of the Forty-Seventh Annual ACM Symposium on Theory of Computing, STOC '15, Association for Computing Machinery, New York, NY, USA, 2015, pp. 117–126, http://dx.doi.org/10.1145/2746539.2746580.
[49] Y. Ding, X. Wu, Y. Meng, Y. Luo, H. Wang, W. Pan, Delving into differentially private transformer, in: Proceedings of the 41st International Conference on Machine Learning, ICML '24, JMLR.org, 2024.
[50] F. Tramer, D. Boneh, Differentially private learning needs better features (or much more data), in: International Conference on Learning Representations, 2021.
[51] D. Yu, H. Zhang, W. Chen, T. Liu, Do not let privacy overbill utility: Gradient embedding perturbation for private learning, in: International Conference on Learning Representations, 2021.
[52] S. De, L. Berrada, J. Hayes, S.L. Smith, B. Balle, Unlocking high-accuracy differentially private image classification through scale, 2022, arXiv:2204.13650.
[53] A. Cheng, J. Wang, X.S. Zhang, Q. Chen, P. Wang, J. Cheng, Dpnas: Neural architecture search for deep learning with differential privacy, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 36, 2022, pp. 6358–6366.
[54] L. Berrada, S. De, J.H. Shen, J. Hayes, R. Stanforth, D. Stutz, P. Kohli, S.L. Smith, B. Balle, Unlocking accuracy and fairness in differentially private image classification, 2023.
[55] N. Papernot, T. Steinke, Hyperparameter tuning with renyi differential privacy, in: International Conference on Learning Representations, 2022.
[56] Y. Huang, G. Yang, Y. Bai, H. Dai, Differential privacy protection for support vector machines for nonlinear classification, Secur. Commun. Networks 2022 (2022) http://dx.doi.org/10.1155/2022/7941915.
[57] Q. Li, Z. Wu, Z. Wen, B. He, Privacy-preserving gradient boosting decision trees, 2019, CoRR abs/1911.04209.