6-23-2022

# Assessing the Reidentification Risks Posed by Deep Learning Algorithms Applied to ECG Data

Arin Ghazarian
*Chapman University*, ghazarian@chapman.edu

Jianwei Zheng
*Chapman University*, zheng120@mail.chapman.edu

Daniele Struppa
*Chapman University*, struppa@chapman.edu

Cyril Rakovski
*Chapman University*, rakovski@chapman.edu



## Recommended Citation

A. Ghazarian, J. Zheng, D. Struppa and C. Rakovski, "Assessing the Reidentification Risks Posed by Deep Learning Algorithms Applied to ECG Data," in *IEEE Access*, vol. 10, pp. 68711-68723, 2022, https://doi.org/10.1109/ACCESS.2022.3185615.

# Assessing the Reidentification Risks Posed by Deep Learning Algorithms Applied to ECG Data

IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 23 May 2022, accepted 14 June 2022, date of publication 23 June 2022, date of current version 1 July 2022.

*Digital Object Identifier 10.1109/ACCESS.2022.3185615*

# Assessing the Reidentification Risks Posed by Deep Learning Algorithms Applied to ECG Data

**ARIN GHAZARIAN, (Member, IEEE), JIANWEI ZHENG, DANIELE STRUPPA, AND CYRIL RAKOVSKI**

Schmid College of Science and Technology, Chapman University, Orange, CA 92866, USA

Corresponding author: Jianwei Zheng (zheng120@mail.chapman.edu)

**ABSTRACT** ECG (Electrocardiogram) data analysis is one of the most widely used and important tools in cardiology diagnostics. In recent years the development of advanced deep learning techniques and GPU hardware have made it possible to train neural network models that attain exceptionally high levels of accuracy in complex tasks such as heart disease diagnoses and treatments. We investigate the use of ECGs as biometrics in human identification systems by implementing state-of-the-art deep learning models. We train convolutional neural network models on approximately 81k patients from the US, Germany and China. Currently, this is the largest research project on ECG identification. Our models achieved an overall accuracy of 95.69%. Furthermore, we assessed the accuracy of our ECG identification model for distinct groups of patients with particular heart conditions and combinations of such conditions. For example, we observed that the identification accuracy was the highest (99.7%) for patients with both ST changes and supraventricular tachycardia. We also found that the identification rate was the lowest for patients diagnosed with both atrial fibrillation and complete right bundle branch block (49%). We discuss the implications of these findings regarding the reidentification risks of patients based on ECG data and how seemingly anonymized ECG datasets can cause privacy concerns for the patients.

**INDEX TERMS** Biometrics, convolutional neural networks (CNN), deep learning, electrocardiogram (ECG), ECG identification, privacy, reidentification.

## I. INTRODUCTION

An Electrocardiogram (ECG) is a recording of the bioelectrical activity of the heart collected from the human body surface. Figure 1 shows an ECG during one normal heartbeat, consisting of several features including the P-wave, QRS complex, T-wave, PR interval, QT interval, PR segment and ST segment. The amplitudes, time intervals and other morphological features in different sections of the ECG signal are used for diagnoses of cardiac conditions, making ECG an important noninvasive tool for detection of heart abnormalities. Arrhythmia is a group of conditions in which the heartbeat has an irregular rate or rhythm. For example, Figure 2 depicts the changes in ECG caused by atrial fibrillation (AFIB), the most common type of arrhythmia that is associated with stroke and heart failure and affects approximately 3% of the US population.

The shape of the ECG signal is unique for each person. ECG is a universal biometric marker since it is present in alive humans and is continuous meaning that we can always capture the ECG from an individual. In the absence of major cardiac events or conditions, an individual's ECG stays relatively unchanged over time, making it possible to create high-accuracy ECG identification systems on multisession ECG data. These qualities make ECG a good candidate as a biometric identifier (electrophysiologic) that can be used for human authentication and identification purposes, similar to fingerprint and iris [1], [2]. Recently, off-the-person ECG acquisition methods have gained more attention in ECG biometric systems since these devices do not need to be attached to the body and are noninvasive. For example, Apple Watch series 6 is capable of capturing a single lead ECG. In addition to being a unique and stable identifier, ECG eliminates the

The associate editor coordinating the review of this manuscript and approving it for publication was Adam Czajka.

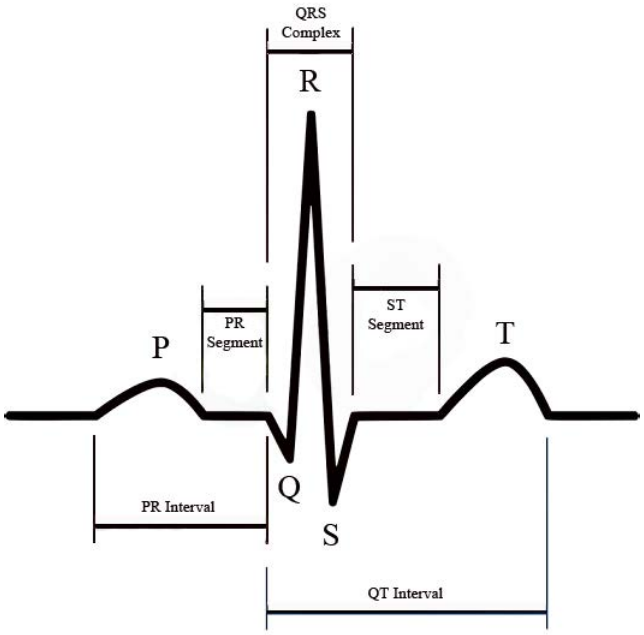


**FIGURE 1.** ECG waveform and segments in lead II for a normal cardiac cycle.

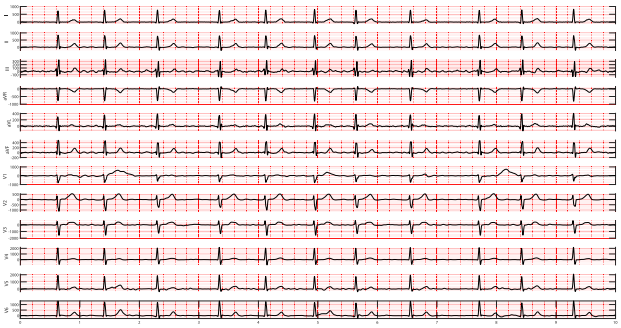

**FIGURE 2.** Twelve-lead ECG showing atrial fibrillation rhythm with no visible P waves that are replaced by coarse fibrillatory waves and an irregularly irregular QRS complex.

aliveness test required in some other forms of biometric systems since the heart signal is an inherently alive biometric. ECG signals are difficult to counterfeit, which makes them even more desirable as a biometric identifier.

The application of ECG as a biometric can be a useful technology, but it also raises serious concerns regarding the potential privacy leakages. For example, an ECG-based biometric system can diagnose and store the heart conditions of users without the patient's consent. Privacy protection becomes even more critical as we collect and store a constantly increasing volume of data from citizens. For example, the emerging medical wearable device technologies capture and store a continuous stream of sensitive health data.

Applications of machine learning techniques for automated analysis of ECG data have been the focus of many recent cardiac research efforts such as arrhythmia classification [3] and accurate prediction of ventricular arrhythmia origins [4]. ECG data have also been used for emotion recognition applications [5], [6]. Automated ECG analysis systems using machine learning techniques typically have multiple steps such as denoising and baseline correction, heartbeat segmentation and QRS detection, feature extraction and model training. Both temporal/morphological features such as QRS duration or amplitude of the P-wave, and frequency domain features such as Fourier or wavelet transformation coefficients have been used by researchers. A variety of machine learning techniques including support vector machines (SVMs), naive Bayes, random forest and neural networks have been used in ECG research. In the recent years, neural networks have been the preferred method for high-accuracy ECG analysis due to the advancements in deep learning algorithms and the availability of fast processors such as GPUs.

In this study, we report our approach in creating a high-accuracy ECG identification system and the implications of such systems for the patients' privacy. The main contributions of this paper are summarized as follows:

- We report the results from our deep learning-based ECG identification system on the largest number of subjects reported in the literature so far (around 81k).
- For the first time, we also assess the accuracy of an ECG identification model for distinct groups of patients with particular heart conditions and combinations of such conditions.
- In a novel analysis, we discuss different reidentification scenarios regarding ECG datasets and provide probability estimations for the reidentification risks in some scenarios.

The results from our research show that ECG is more sensitive health care information than previously thought and needs to be protected by the privacy laws and regulations. For instance, we recommend that ECG data be included in the list of HIPAA identifiers.

## II. LITERATURE REVIEW

The idea of leveraging ECG as a biometric identifier was introduced by Forsen *et al.* in 1977 [7]. Biel *et al.* [1] were the first to implement an ECG biometric system in 1999. Although, subsequent studies have reported high accuracies in ECG identification, all of them were based on a small number of subjects ranging from ten to a couple of hundreds. This undermines the results since ECG identification systems in real-world scenarios are supposed to run on a large population. In this section, we will review the ECG identification literature, focusing on the number of subjects used in each study and the accuracy achieved. Labati *et al.* [8] extracted features from ECG using a CNN-based deep learning model and achieved 100% accuracy on approximately 50 human subjects. Belo *et al.* [9] leveraged a temporal convolutional neural network (TCNN) and recurrent neural network (RNN) for both ECG identification and

authentication. Overall, the TCNN model outperformed the RNN achieving 100%, 96% and 90% accuracy on the Fantasia (40 subjects), MIT-BIH (47 subject), and CYBHi (63 subjects) databases, respectively. Salloum and Kuo [10] used recurrent neural networks (RNNs) with long short-term memory (LSTM) and gated recurrent units (GRUs), reaching a 100% identification rate on 90 subjects from the public ECG-ID database. Deshmane and Madhe [11] proposed a CNN-based approach that achieved 81.33%, 96.95%, 94.73% and 92.85% accuracies on the MITBIH (47 subjects), FANTASIA (40 subjects), NSRDB (18 subjects) and QT databases (105 subjects), respectively. Eduardo *et al.* [12] used autoencoders for denoising and feature extraction in an ECG biometric system. Zhang *et al.* [13] achieved an average identification rate of 93.5% using a multiresolution CNN on datasets of 18 to 47 subjects. Li *et al.* [14] implemented two cascaded CNNs: one for feature extraction from ECG heartbeats and another for identification. They achieved 99.52% accuracy on 184 subjects.

In general, deep learning models and CNNs have been used for different tasks in automated ECG analysis. Li *et al.* [15] proposed a new technique that combines convolutional neural networks and distance distribution matrices (DDM) to classify congestive heart failure patients from normal subjects in the MIT-BIH dataset. DDMs were used in entropy calculations, and CNN models such as S_Inception_v4 were used to learn the pattern of the data distributions hidden in the generated distribution matrices. They achieved an accuracy of 81.85. Oh *et al.* [16] proposed a method leveraging a combination of a convolutional neural network (CNN) and long short-term memory (LSTM) for the diagnosis of normal sinus rhythm, left bundle branch block (LBBB), right bundle branch block (RBBB), atrial premature beats (APB) and premature ventricular contraction (PVC) on ECG signals. They used variable length segments from the MIT-BIT arrhythmia dataset, achieving an accuracy of 98.10%, sensitivity of 97.50% and specificity of 98.70%.

Limited research has been conducted on predicting demographics such as age, sex and race from ECG signals. Attia *et al.* [17] trained a CNN to predict age and sex on 10-second samples of 12-lead ECG signals from 499,727 patients. They achieved an accuracy of 90.4% for sex classification and an average error of 6.9±5.6 years for age estimation on a separate validation set of 275,056 patients. Khan *et al.* [18] trained a sex classification model with an accuracy of 95.2%. Cabra *et al.* [19] reported 94% accuracy in automated classification of sex using ECG samples. Wiggins *et al.* [20] presented a genetically evolved Bayesian classifier for age detection capable of assigning patients into young and elderly groups with an AUC of 86.25%. Noseworthy *et al.* [21] built neural network models to discern racial subgroups including African American, White, Hispanic/Latino, Asian and Indigenous People or Alaskan Native with an accuracy of 56.2%.

## III. DATA

The 12-lead ECG data used in this work consisted of four open-access research resources and a new open dataset from the Ningbo First Hospital. The first open access dataset [22] was used in the China Physiological Signal Challenge in 2018. This source contains 10,330 ECGs. Each recording is between 6 and 144 seconds long with a sampling frequency of 500 Hz. The second source is the Physikalisch-Technische Bundesanstalt (PTB) ECG dataset [23], which consists of 21,837 clinical 12-lead ECGs from 18,885 patients of 10 seconds in length. The raw waveform data were annotated by up to two cardiologists, who potentially assigned multiple ECG statements to each record. The third source is a Georgia database [24] that encompassed 10,334 ECGs and represented a unique demographic of the southeastern United States. Each recording is between 5 and 10 seconds long with a sampling frequency of 500 Hz. The fourth database [25] contains 12-lead ECGs from 10,646 patients with a 500-Hz sampling rate that features 11 common rhythms and 67 additional cardiovascular conditions all labeled by professional experts. This dataset consists of 10-second, 12-dimensional ECGs and labels for rhythms and other conditions for each subject. In addition, a new dataset from the Ningbo First Hospital, including 34,320 ECG recordings, was collected for study. The institutional review board of Ningbo First Hospital approved this study and waived the requirement to obtain informed consent. Cardiologist-supervised physicians interpreted each recording and gave cardiac condition labels and ECG findings. Finally, there were 88 cardiac conditions present in the combined data that contained 87,467 ECG recordings.

## IV. PREPROCESSING

To improve the data input quality supplied to the neural network, we carried out a three-stage noise reduction process including a Butterworth low-pass filter to remove high-frequency noise (above 50 Hz), Robust LOESS to eliminate baseline wandering, and Nonlocal Means (NLM) to remove residual noise [3]. These filters help to reduce the noise in ECG signals caused by known sources such as power line interference, electrode contact noise, motion artifacts, skeletal muscle contraction and random noise. The low-frequency ($<$0.5 Hz) baseline wandering noise component could be caused by respiration. Power line interference is the major cause of the high-frequency (50-60 Hz) noise component. R-peak to R-peaks are extracted from each ECG recording and downsampled to 300 points.

## V. ECG IDENTIFICATION MODEL

As shown in Figure 3, we implemented a convolutional neural network with six one-dimensional convolutional sequences, of which the first five layers (i.e. layers 1-1, 1-2, 2-1, 2-2 and 3) are horizontal (temporal) convolutions of kernel size $1 \times 5$ and the sixth layer (layer 4) is a vertical (spatial) convolution of kernel size $12 \times 1$. Each convolution is followed by

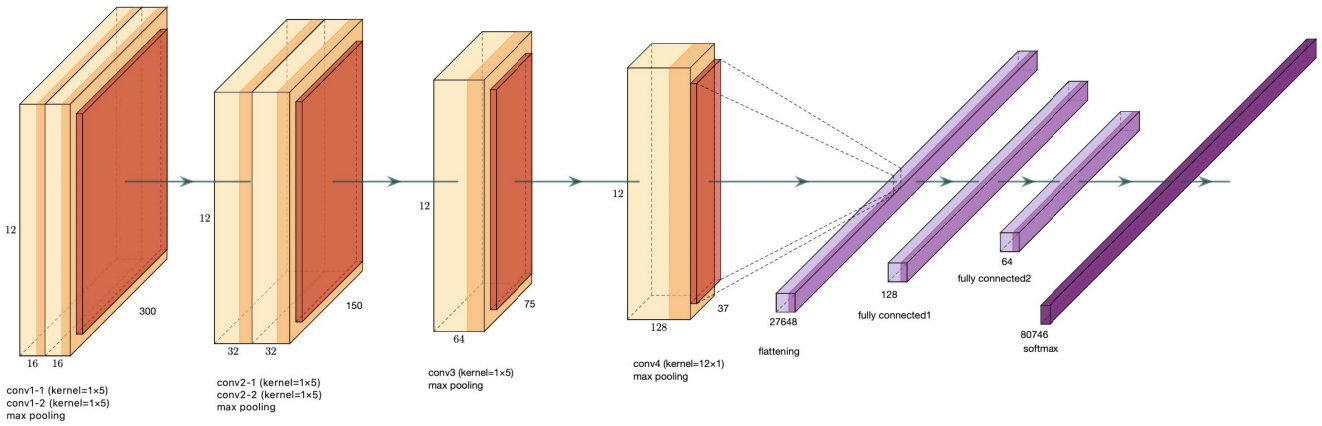


**FIGURE 3.** Architecture of the deep learning model for ECG identification.

batch normalization, and there is a max-pooling layer after convolutional layers 1-2, 2-2, 3 and 4 with a pool size of $1 \times 2$. The convolutional layers are followed by a flattening layer, and two sequences of dense and dropout layers. The final layer consists of softmax units equal to the number of patients in the dataset. We used relu activation function for all of the convolutional and dense layers. Overall, the architecture resembles a temporal-spatial model where the first five convolutional sequences try to discover patterns across each of the 12 ECG leads and the last sequence looks for patterns across all leads at one point in time. The input vector is a $12 \times 300$ vector representing the 12-lead data consisting of R-to-R interval measurements. The labels are the encoded IDs for the patient in the database. The Adam optimizer is used with a sparse categorical cross entropy loss function. We tuned the model parameters by trying different values for a set of different parameters and choosing the best parameter. For example, we tried different ranges of values for parameters such as dropout rate, size of convolution windows, number of neurons, activation function, optimizer, learning rate and batch size.

### A. OVERALL IDENTIFICATION RATE

An intrapatient validation method was adopted, as it is a fair evaluation method for identification systems. We randomly selected 20% of heartbeats from each patient to be used in the validation set, and the rest of the heartbeats were used in the training set. Thus, all patients were present in both the training and validation sets but with distinct, nonoverlapping heartbeat data. There were a total of 1,142,859 R-to-R interval data from 80,746 patients split into two sets of sizes (914,287 and 228,572) for the training and validation samples, respectively. Figure 4 shows the good fit in the convergence curve from the model training. Our optimal CNN model attained an overall accuracy of 95.69% over all R-to-R interval validation data for the entire sample of 80,746 patients.

The deep learning model employed to reidentify subjects based on their ECG recordings needs to possess exceptional inherent quality, especially as the number of patients in the data increases. A large sample of subjects simultaneously provides two challenges, increases the probability of observing subgroups of similar ECG profiles and dramatically increases the number of possible incorrect identities for any accuracy level. For example, given a sample size of $n$, the proportion of identification selections that entail an accuracy of $p100\%$ is:

$$\frac{\binom{n}{np}[\sum_{i=1}^{n-np}(-1)^{i+1}1/i!](n-np)!}{n!}. \tag{1}$$

Equation (1) can be approximated and simplified:

$$\frac{1}{(np)!e}. \tag{2}$$

Last, using Stirling's approximation Equation (2) yields:

$$\frac{e^{np-1}}{\sqrt{2\pi np}(np)^{np}}. \tag{3}$$

In our study, with a sample size $n$ of 80,746, we attained an accuracy of 95.69%. Formula (3) shows that the proportion of such favorable reidentification selection is practically zero.

### B. IDENTIFICATION RATE PER CONDITION

To assess the privacy risks for cardiology patients posed by ECG identification systems, we calculated the identification rates for patients suffering from several heart conditions. Obviously, for a specific group of patients, the higher the misidentification rate, the lower the privacy risk. In the discussion section, we explain how ECG identification can be used to reidentify patients from anonymized datasets.

In our per-condition analysis, we considered single- and multiple-condition scenarios. In the single-condition analysis, a misidentified patient with one or several conditions was counted as misidentified for all conditions. The total misidentification rates per condition were calculated as the proportion of misidentified patients with each condition among all patients with the same condition. In the multiple-condition analysis, we considered all diagnoses of a patient as a single

complex category. This allowed us to assess the identification rates for patients suffering from a particular combination of conditions. The identification accuracies per all single conditions are shown in Table 5. The identification rates for all joint conditions with more than one hundred samples are listed in Table 6. Note that in the joint condition analysis results table, we have rows representing a single condition since some patients had only a single diagnosis in the original data. Table 7 shows the full diagnosis names for the condition codes in Tables 5 and 6.

Based on Table 5, some single conditions, such as ventricular fibrillation, sinus arrest, left bundle branch block, myocardial infarction and premature ventricular contractions had low identification rates ( 54.96%,66.67%,83.43%, 85.14% and 86.04%, respectively), while healthy sinus rhythm and conditions such as shortened PR interval, supraventricular tachycardia or counterclockwise vector cardiographic loop had high identification rates (98.15%, 100%, 99.55% and 98.76%, respectively). Physiologically speaking, one potential hypothesis could be since the ECG morphologies of ventricular fibrillation, sinus arrest, left bundle branch block, myocardial infarction and premature ventricular contractions among patients are similar and do not have many individualized features, the identification rate will be lower than normal rhythm. Based on the joint condition analysis (Table 6), patients with both atrial fibrillation and complete right bundle branch block had a very low identification rate (48.95%), making it almost impossible to identify these groups of patients based on their ECG. On the other hand, patients with both ST changes and supraventricular tachycardia had a very high identification rate (99.7%), putting these patients in a very high-risk group regarding privacy. One observation is that in contrast to common expectations, some conditions such as patients with both ST changes and supraventricular tachycardia or patients diagnosed with sinus tachycardia had an even better identification rate than normal sinus rhythm. Finally, for patients with pacemakers, the identification rate was 91.11% in the joint list (i.e. patients with pacemakers and no other condition) and 89.29% in the single disease analysis (i.e. patients with pacemakers and other potential conditions).

For example, consider the scenario where a patient diagnosed with both ST changes and supraventricular tachycardia contributes to two different research datasets. Both datasets are fully anonymized. One of them contains an ECG sample, age and sex, and the other contains an ECG sample and zip code. If we join these two datasets using an ECG identification system, then we can find the subjects who appear in both, giving us age, sex and zip code for the patients who appear in both. The uniqueness for this combination in the United States is 0.04%. We also know that the average identification rate for patients diagnosed with both ST changes and supraventricular tachycardia is 0.997. Multiplying these two numbers (under independence) gives us the probability of identifying this patient uniquely and correctly: $0.997 \times 0.0004 \approx 0.0004$. Now, if we have birth date instead of age, then this probability increases to $0.997 \times 0.871 \approx 0.8684$.

**TABLE 1. Identification rate per sex.**

| Sex | Number of Samples | Identification Rate |
|---|---|---|
| male | 120,413 | 95.13% |
| female | 104,818 | 96.59% |

**TABLE 2. Identification rate per sex for sinus rhythms only.**

| Sex | Number of Samples | Identification Rate |
|---|---|---|
| male | 15,773 | 98.66% |
| female | 20,699 | 99.18% |

**TABLE 3. Uniqueness of US population [27].**

| Identifiers | Uniqueness |
|---|---|
| sex, age, 5-digit ZIP | 0.04% |
| sex, date of birth, ZIP | 87.1% |
| sex, date of birth, county | 18.1% |
| sex, 2year age range, place | 0.01% |
| sex, Year of birth, county | 0.00004% |

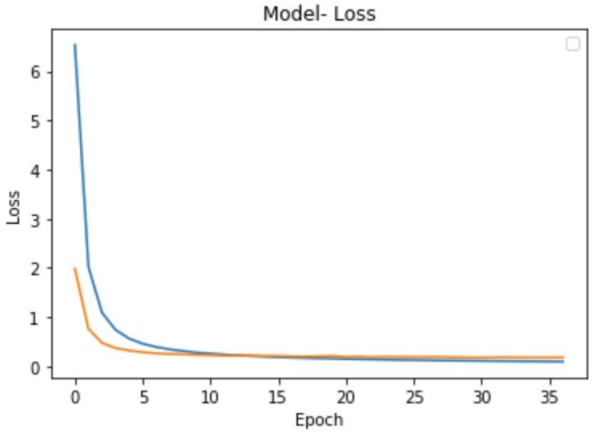


**FIGURE 4. Training loss convergence curve.**

### C. IDENTIFICATION RATE PER SEX AND AGE

We also assessed the identification rates across sexes and age groups. There were 2,996 patients with unknown sex and 649 with unknown age in the data, which were removed from this analysis. Table 1 and Figure 5 summarize our results, which indicate that males had a slightly lower identification rate than females and that older patients had a lower identification rate than younger patients. This implies higher privacy risks for female or younger subjects.

We performed a similar analysis on the subset of patients with healthy normal sinus rhythms to investigate the presence of differences in identification accuracies across sexes or age groups among healthy individuals. The results are shown in Table 2 and Figure 6. Again, male patients had a lower identification rate than females, and older patients had lower identification rates than the younger subjects.

## VI. DISCUSSION

We designed and implemented a deep learning model capable of identifying subjects based on 12-lead ECG data. Previous

**TABLE 4.** HIPAA 18 identifiers.

| Number | Identifier | Notes |
|---|---|---|
| 1 | Name | |
| 2 | Address | all geographic subdivisions smaller than state, including street address, city county, and zip code |
| 3 | Dates | All elements (except years) of dates related to an individual including birthdate, admission date, appointments date, payments date, discharge date, date of death, and exact age if over 89) |
| 4 | Telephone number | |
| 5 | Fax number | |
| 6 | Email address | |
| 7 | Social Security number | |
| 8 | Medical record number | |
| 9 | Health plan/insurance beneficiary number | |
| 10 | Account number | |
| 11 | Certificate / license number | |
| 12 | Any vehicle identifiers (e.g. license plate number) | |
| 13 | Device identifiers and serial numbers | |
| 14 | Web URLs (Links) | |
| 15 | Internet Protocol (IP) address | |
| 16 | Biometric identifiers (finger / retinal / voice) | |
| 17 | Photographic images | Photographic images are not limited to images of the face |
| 18 | Any other characteristic that may be used to uniquely identify an individual | |

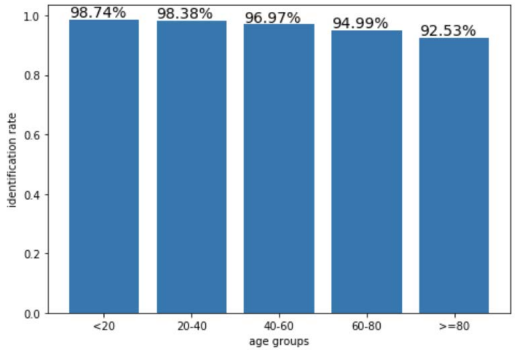


**FIGURE 5.** Identification rate per age group.

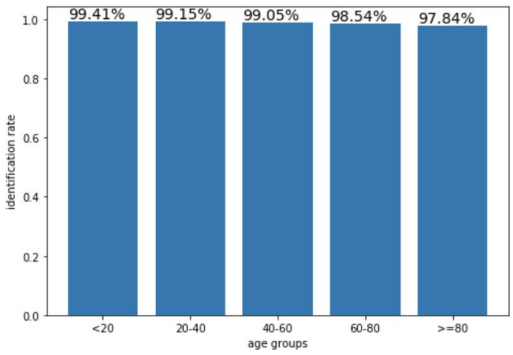


**FIGURE 6.** Identification rate per age group for sinus rhythms only.

results had severe limitations due to limited sample sizes ranging from a couple of dozen to approximately one thousand subjects. In this study, we trained a model with the largest number of subjects: 80,746. We attained an exceptionally high overall accuracy of 95.69%. In the following subsections, we discuss the implications of our findings for the privacy of patients.

### A. RE-IDENTIFICATION RISKS DUE TO DEEP LEARNING-BASED ECG ANALYSIS

ECG identification algorithms create the potential for reidentification of individuals in ECG databases. Reidentification is the practice of discovering the identity of individuals in an anonymized database by matching the records with publicly available information (auxiliary data). There is an

**TABLE 5.** Identification rate per single condition.

| Condition Name | Number of Heartbeats in Validation | Identification Rate |
|---|---|---|
| AAR | 39.0 | 100.0 |
| SPRI | 66.0 | 100.0 |
| PMI | 52.0 | 100.0 |
| AJR | 74.0 | 100.0 |
| IIAVBII | 7.0 | 100.0 |
| PTW | 11.0 | 100.0 |
| LAHV | 4.0 | 100.0 |
| Brugada | 13.0 | 100.0 |
| SVT | 3552.0 | 99.55 |
| AMI | 216.0 | 99.54 |
| RAHV | 120.0 | 99.17 |
| JTach | 98.0 | 98.98 |
| TPW | 857.0 | 98.95 |
| CCVCL | 1371.0 | 98.76 |
| TTW | 892.0 | 98.65 |
| Atrial Rhythm | 417.0 | 98.56 |
| PPW | 319.0 | 98.43 |
| SB | 29959.0 | 98.42 |
| RRWP | 121.0 | 98.35 |
| ERe | 1048.0 | 98.19 |
| SNR | 62107.0 | 98.15 |
| STach | 29104.0 | 97.96 |
| AVB | 477.0 | 97.9 |
| RAD | 2673.0 | 97.68 |
| LVHV | 11611.0 | 97.66 |
| LQRSV | 4038.0 | 97.62 |
| TAb | 25232.0 | 97.21 |
| LPFB | 521.0 | 97.12 |
| TInv | 10709.0 | 97.09 |
| LPR | 847.0 | 97.05 |
| LMI | 572.0 | 97.03 |
| Intraventricular Block | 1159.0 | 96.98 |
| CVCL | 944.0 | 96.93 |
| SA | 7780.0 | 96.83 |
| AFL | 7451.0 | 96.82 |
| Atrial Escape Beat | 31.0 | 96.77 |
| LAD | 5857.0 | 96.76 |
| LQT | 4342.0 | 96.75 |
| LVH | 11037.0 | 96.62 |
| PRWP | 1996.0 | 96.59 |
| QAb | 3566.0 | 96.55 |
| WPW | 402.0 | 96.52 |
| LAnFB | 5149.0 | 96.5 |
| LAH | 3246.0 | 96.4 |
| RBBB | 1883.0 | 96.18 |
| PWC | 562.0 | 96.09 |
| VFL | 24.0 | 95.83 |
| STC | 46469.0 | 95.63 |
| RVH | 1068.0 | 95.32 |
| CLBBB | 2018.0 | 95.24 |
| UAb | 271.0 | 95.2 |

**TABLE 5.** Identification rate per single condition.

| Condition Name | Number of Heartbeats in Validation | Misidentification Rate |
|---|---|---|
| AVD | 102.0 | 95.1 |
| AVC | 4453.0 | 94.32 |
| AnMI | 7423.0 | 94.3 |
| InMI | 7458.0 | 94.25 |
| RAH | 500.0 | 94.2 |
| IRBBB | 4682.0 | 94.02 |
| IAVB | 9410.0 | 94.0 |
| STD | 12546.0 | 93.51 |
| PVT | 532.0 | 93.23 |
| ATach | 1199.0 | 92.99 |
| RAVC | 47.0 | 91.49 |
| Brady | 1182.0 | 91.2 |
| JE | 153.0 | 90.85 |
| AF | 34954.0 | 90.72 |
| ILBBB | 677.0 | 90.25 |
| PAC | 9091.0 | 90.11 |
| BPAC | 171.0 | 90.06 |
| IIAVB | 236.0 | 89.83 |
| CAVB | 230.0 | 89.57 |
| STE | 4189.0 | 89.5 |
| PR | 3128.0 | 89.29 |
| IIAVBI | 65.0 | 89.23 |
| FB | 261.0 | 88.89 |
| CRBBB | 14719.0 | 87.6 |
| SPVB | 487.0 | 87.47 |
| AIVR | 23.0 | 86.96 |
| OldMI | 5809.0 | 86.64 |
| PVC | 13206.0 | 86.04 |
| MI | 1555.0 | 85.14 |
| JPC | 47.0 | 85.11 |
| VEsR | 69.0 | 84.06 |
| LBBB | 2142.0 | 83.43 |
| AVJR | 288.0 | 82.64 |
| VEsB | 78.0 | 76.92 |
| SAB | 72.0 | 72.22 |
| SARR | 30.0 | 66.67 |
| VF | 131.0 | 54.96 |

incorrect historical belief that anonymizing data obtained by removing identifiers from a dataset protects the privacy of subjects. There have been many instances of reidentification of individuals in anonymized research datasets. For example, in 1997, a researcher from MIT reidentified the governor of Massachusetts in an anonymized research health care dataset [26] by matching these data with the publicly available voter registration data.

A potential reidentification scenario in ECG datasets could occur when an individual contributes data to two or more different research datasets. For example, database A has an ECG sample, sex and date of birth, while database B has an ECG sample and zip code. By simply matching the ECG columns in both databases using an identification system, one can discover the individuals who appear in both datasets and obtain a complete profile of the individual by joining their records. In this case, we have sex and date of birth from database A and zip code from database B. These three demographic attributes might be enough to uniquely identify someone, as 87% of US citizens can be uniquely identified only by having their date of birth, sex and zip code [27], [28]. While each of these demographics alone (quasi-identifiers) is insufficient to identify someone, their combination can be unique for a considerable percentage of the population.

**TABLE 6.** Identification rate considering joint conditions.

| Condition Name | Number of Heartbeats in Validation | Identification rate |
|---|---|---|
| STC SVT | 1003.0 | 99.7 |
| STach | 9548.0 | 99.53 |
| LVHV SB | 1979.0 | 99.34 |
| STach TAb | 1477.0 | 99.26 |
| SB | 15009.0 | 99.05 |
| SB TAb | 1124.0 | 99.02 |
| SNR | 37277.0 | 98.88 |
| SNR TAb | 2565.0 | 98.83 |
| SNR STC | 1282.0 | 98.52 |
| STC STach | 2870.0 | 98.5 |
| SA | 3524.0 | 98.07 |
| STC | 7394.0 | 97.57 |
| AFL | 1230.0 | 97.48 |
| OldMI | 2092.0 | 97.32 |
| InMI SNR | 1843.0 | 96.69 |
| STD | 4720.0 | 96.69 |
| AF TAb | 1908.0 | 96.65 |
| CRBBB | 7478.0 | 96.47 |
| IAVB | 3160.0 | 95.28 |
| AF | 10557.0 | 93.57 |
| AF STC | 2294.0 | 93.5 |
| PVC | 4904.0 | 92.48 |
| PR | 1136.0 | 91.11 |
| PAC | 2785.0 | 90.59 |
| STE | 1196.0 | 83.03 |
| AF CRBBB | 1283.0 | 48.95 |

Table 3 shows how different combinations of quasi-identifiers such as sex, birth date, age and race can uniquely identify citizens of the United States [27]. The table reports average numbers across the United States. However, some specific geographic regions have much higher uniqueness rates for these identifiers. Additionally, rare scenarios such as patients over 90 years old or a very small population of an ethnic/race group living in a zip code can make them more vulnerable to reidentification attacks.

### B. RE-IDENTIFICATION RISKS DUE TO ECG-BASED DEMOGRAPHICS PREDICTION

There have been reports regarding high accuracies of ECG-based age and sex detection models [17]. Therefore, we can extract the age and sex of the patient from the sample. Typically, patients live in the proximity of the hospital or clinic where their ECG was captured. Consequently, if we know the hospital where the ECG sample was recorded, then we can assume that the patient lives in that zip code, city or county with some probability. Thus, by just having a sample ECG and knowing the hospital where it was captured, we can know the age, sex and zip code of the patient with some probabilities. The combination of these three elements (age, sex and zip code) might be enough to fully identify the patient and locate their residence via online people search databases and auxiliary public database searches. In this scenario, we might be able to identify and locate a patient merely using an ECG sample and no other information.

Although we cannot estimate the exact value for the birth date (year, month and day) based on ECG deep learning analysis, we can precisely estimate the age. If our database has the date for capturing the ECG from the patient, then we can reference the value of age to that date and calculate the value for the year of birth. Year of birth is 365 times less identifiable than an exact date of birth but is more informative in terms of uniqueness than age. Additionally, race might be detectable from an ECG sample, which can significantly narrow down the search for the patient.

### C. IMPLICATIONS FOR PRIVACY LAWS

There are different opinions on the criminalization of wrongful reidentification. Some advocates of reidentification criminalization believe that doing so will eventually have a great impact on health and medical discoveries through big data analysis since clear laws and regulations will ease the data collection and analysis [29]–[31]. There are also opponents who believe that fewer restrictions in data sharing will revolutionize medical research [32].

**TABLE 7.** Diagnosis codes mapping table.

| Diagnosis Code | Full Diagnosis Name |
|---|---|
| SR | sinus rhythm |
| SRNORM | sinus normal rhythm |
| NORM | normal ECG |
| AAR | accelerated atrial rhythm |
| AF | atrial fibrillation |
| AFL | atrial flutter |
| AIVR | accelerated idioventricular rhythm |
| AJR | accelerated junctional rhythm |
| AMI | acute myocardial infarction |
| ATach | atrial tachycardia |
| AVB | av block |
| AVC | Aberrant Ventricular Conduction |
| AVD | auriculoventricular dissociation |
| AVJR | atrioventricular junctional rhythm |
| AnMI | anterior myocardial infarction |
| Atrial Escape Beat | Atrial Escape Beat |
| Atrial Rhythm | Atrial Rhythm |
| BPAC | blocked premature atrial contraction |
| Brady | bradycardia |
| Brugada | Brugada |
| CAVB | complete atrioventricular block |
| CCVCL | Counterclockwise vectorcardiographic loop |
| CLBBB | complete left bundle branch block |
| CRBBB | complete right bundle branch block |
| CVCL | Clockwise vectorcardiographic loop |
| ERe | early repolarization |
| FB | fusion beats |
| IAVB | 1st degree av block |
| IIAVB | 2nd degree av block |
| IIAVBI | Mobitz type I second degree atrioventricular block |
| IIAVBII | Mobitz type II atrioventricular block |
| ILBBB | incomplete left bundle branch block |
| IRBBB | incomplete right bundle branch block |
| InMI | inferior myocardial infarction |
| Intraventricular Block | Intraventricular Block |
| JE | junctional escape |
| JPC | junctional premature complex |
| JTach | junctional tachycardia |
| LAD | left axis deviation |
| LAH | left atrial hypertrophy |
| LAHV | Left atrial high voltage |
| LAnFB | left anterior fascicular block |
| LBBB | left bundle branch block |
| LMI | lateral myocardial infarction |
| LPFB | left posterior fascicular block |
| LPR | prolonged pr interval |
| LQRSV | low qrs voltages |
| LQT | prolonged qt interval |
| LVH | left ventricular hypertrophy |
| LVHV | left ventricular high voltage |
| MI | myocardial infarction |

**TABLE 7.** *(Continued.)* **Diagnosis codes mapping table.**

| Diagnosis Code | Full Diagnosis Name |
|---|---|
| PAC | premature atrial contraction |
| PMI | Posterior myocardial infarction |
| PPW | Prolonged P wave |
| PR | pacing rhythm |
| PRWP | Poor R wave Progression |
| PTW | Prolonged T wave |
| PVC | premature ventricular contractions |
| PVT | paroxysmal ventricular tachycardia |
| PWC | P wave change |
| QAb | qwave abnormal |
| RAD | right axis deviation |
| RAH | right atrial hypertrophy |
| RAHV | right atrial high voltage |
| RAVC | retrograde atrioventricular conduction |
| RBBB | right bundle branch block |
| RRWP | reversed R wave progression |
| RVH | right ventricular hypertrophy |
| SA | sinus arrhythmia |
| SAB | sinoatrial block |
| SARR | sinus arrest |
| SB | sinus bradycardia |
| SNR | sinus rhythm |
| SPRI | shortened pr interval |
| SPVB | supraventricular premature beats |
| SQT | decreased qt interval |
| STC | st changes |
| STD | st depression |
| STE | st elevation |
| STach | sinus tachycardia |
| SVT | supraventricular tachycardia |
| TAb | t wave abnormal |
| TInv | t wave inversion |
| TPW | Tall P wave |
| TTW | Tall tented T wave |
| UAb | u wave abnormal |
| VEsB | ventricular escape beat |
| VEsR | ventricular escape rhythm |
| VF | ventricular fibrillation |
| VFL | ventricular flutter |
| WPW | wolff parkinson white pattern |
| OldMI | old myocardial infarction |

The Health Insurance Portability and Accountability Act (HIPAA or the Kennedy–Kassebaum Act) of 1996 is a United States federal statute that establishes regulatory thresholds regarding data from health records to protect all individually identifiable health information from patients. Similarly, in the European Union, the General Data Protection Regulation (GDPR) is a data protection and privacy regulation that sets guidelines for the collection and processing of all types of personal information (including medical and health care) for citizens. It gives individuals control over their personal data. For instance, GDPR-compliant pseudonymization is a set of guidelines to reduce privacy risks by assuring that data cannot be attributed to a specific subject without the use of separately kept additional information.

The HIPAA Safe Harbor standard states that a dataset derived from health records should not contain any of the 18

HIPAA identifiers to be considered deidentified [33]. Table 4 lists the 18 identifiers stated by HIPAA. Some of these identifiers are explicit, such as address, phone number or email, which could be used to directly contact the patient [27]. Even though identifier number 16 is for biometric identifiers, ECG is not explicitly mentioned. In addition to being a biometric identifier, ECG can also potentially reveal other information about patients such as age, sex, race and health conditions. As shown in the previous subsections, this information can be used to identify the person and even find their explicit identifiers such as address, phone number and name, which are HIPAA identifiers 1, 2 and 4. Benitez and Malin [34] estimated that the percentage of a state's population to be vulnerable to unique reidentification attacks even when HIPAA Safe Harbor is applied ranges from 0.01% to 0.25%. The findings from our research indicate that ECG is more sensitive health information than previously thought and requires more attention from a privacy perspective. The results from our research suggest that ECG should be added to the list of HIPAA identifiers as the HIPPA 19th identifier; otherwise, it should be clearly mentioned as part of identifier 16.

HIPAA also has regulations regarding the right to access your Protected Health care Information (PHI) including health condition, treatment plan, notes, images, lab results and billing information. Likewise, the European GDPR defines the right of access to personal data and information for subjects. To comply with the GDPR, the data collector must provide a copy of the actual data upon the subject request, i.e., what we have stored about you.

However, data access rights become tricky in regard to data such as ECG, based on which other information such as age, sex and diseases can be extracted on the fly. While a single ECG sample without any other data related to it seems to be harmless with respect to the owner's privacy, it can reveal sensitive derivative information about its owner. In this case, we recommend that in addition to the right to access the actual copy of data, data collectors must provide the subject with enough information on the potential of data privacy infringement associated with that piece of data. For instance, organizations collecting ECG data could provide enough information to patients regarding potential derivative data such as age, sex, race, heart conditions and life habits (such as smoking and level of alcohol consumption).

## VII. CONCLUSION

ECG has the potential to uniquely identify individuals in a large population. Although deep learning models such as convolutional neural networks are difficult to interpret and understand, they are able to learn complex tasks such as ECG identification with high accuracy. This has revolutionized application areas such as ECG as a biometric where interpretability of the model is not a concern. Additionally, we showed that identification accuracy per condition, age and sex differs significantly. These facts imply that patients with different conditions and demographics are exposed to different levels of reidentification risks when contributing data to ECG datasets. As discussed in this paper, the amount of this privacy risk can be quantified in terms of how uniquely and accurately an individual in an ECG dataset can be identified. Patients can use these risk estimations to decide whether they contribute their ECG recording to a research dataset. Our research suggests that ECG is more sensitive information than considered today. Thus, privacy regulations such as HIPAA in the United States or the European GDPR should add explicit guidelines related to the collection and sharing of ECG data to reduce privacy risks for patients.

## VIII. DATA AND CODE AVAILABILITY STATEMENT

The data that support the findings of this study are openly available at the following URLs:

- PTB-XL dataset from https://physionet.org/content/ptb-xl/1.0.1/
- 2018 China Physiological Signal Challenge dataset from http://2018.icbeb.org/Challenge.html
- The Georgia 12-lead ECG Challenge dataset from https://physionetchallenges.org/2020/
- The Ningbo First Hospital 12-lead ECG dataset from https://physionetchallenges.org/2021/

The source code for ECG denoising is available at https://github.com/zheng120/ECGDenoisingTool, and the source code for training the models and the analysis used in this paper is also openly available at https://github.com/arin-gzn/ECG-Identification.

## REFERENCES

[1] L. Biel, O. Pettersson, L. Philipson, and P. Wide, "ECG analysis: A new approach in human identification," *IEEE Trans. Instrum. Meas.*, vol. 50, no. 3, pp. 808–812, Jun. 2001.

[2] C. Carreiras, A. Lourenço, A. Fred, and R. Ferreira, "ECG signals for biometric applications—Are we there yet?" in *Proc. 11th Int. Conf. Informat. Control, Autom. Robot.*, 2014, pp. 765–772.

[3] J. Zheng, H. Chu, D. Struppa, J. Zhang, S. M. Yacoub, H. El-Askary, A. Chang, L. Ehwerhemuepha, I. Abudayyeh, A. Barrett, G. Fu, H. Yao, D. Li, H. Guo, and C. Rakovski, "Optimal multi-stage arrhythmia classification approach," *Sci. Rep.*, vol. 10, no. 1, p. 2898, Dec. 2020. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/32076033

[4] J. Zheng, G. Fu, I. Abudayyeh, M. Yacoub, A. Chang, W. W. Feaster, L. Ehwerhemuepha, H. El-Askary, X. Du, B. He, M. Feng, Y. Yu, B. Wang, J. Liu, H. Yao, H. Chu, and C. Rakovski, "A high-precision machine learning algorithm to classify left and right outflow tract ventricular tachycardia," emphFront. Physiol., vol. 12, Feb. 2021, doi: 10.3389/fphys.2021.641066.

[5] S. Koelstra, C. Muhl, M. Soleymani, J. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras, "DEAP: A database for emotion analysis ;using physiological signals," *IEEE Trans. Affect. Comput.*, vol. 3, no. 1, pp. 18–31, Jan./Mar. 2012.

[6] S. Brás, J. H. T. Ferreira, S. C. Soares, and A. J. Pinho, "Biometric and emotion identification: An ECG compression based method," *Frontiers Psychol.*, vol. 9, p. 467, Apr. 2018, doi: 10.3389/fpsyg.2018.00467.

[7] G. Forsen, M. Nelson, and R. Staron, *Personal Attributes Authentication Techniques*. Fort Belvoir, VA, USA: Defense Technical Information Center, 1977. [Online]. Available: https://books.google.com/books?id=tbs4OAAACAAJ

[8] R. D. Labati, E. Muñoz, V. Piuri, R. Sassi, and F. Scotti, "Deep-ECG: Convolutional neural networks for ECG biometric recognition," *Pattern Recognit. Lett.*, vol. 126, pp. 78–85, Sep. 2019. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167865518301077

[9] D. Belo, N. Bento, H. Silva, A. Fred, and H. Gamboa, "ECG biometrics using deep learning and relative score threshold classification," *Sensors*, vol. 20, no. 15, p. 4078, Jul. 2020. [Online]. Available: https://www.mdpi.com/1424-8220/20/15/4078

[10] R. Salloum and C.-C.-J. Kuo, "ECG-based biometrics using recurrent neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2017, pp. 2062–2066.

[11] M. Deshmane and S. Madhe, "ECG based biometric human identification using convolutional neural network in smart health applications," in *Proc. 4th Int. Conf. Comput. Commun. Control Autom. (ICCUBEA)*, Aug. 2018, pp. 1–6.

[12] A. Eduardo, H. Aidos, and A. Fred, "ECG-based biometrics using a deep autoencoder for feature learning—An empirical study on transferability," in *Proc. 6th Int. Conf. Pattern Recognit. Appl. Methods*, 2017, pp. 463–470.

[13] Q. Zhang, D. Zhou, and X. Zeng, "HeartID: A multiresolution convolutional neural network for ECG-based biometric human identification in smart health applications," *IEEE Access*, vol. 5, pp. 11805–11816, 2017.

[14] Y. Li, Y. Pang, K. Wang, and X. Li, "Toward improving ECG biometric identification using cascaded convolutional neural networks," *Neurocomputing*, vol. 391, pp. 83–95, May 2020. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0925231220300485

[15] Y. Li, Y. Zhang, L. Zhao, Y. Zhang, C. Liu, L. Zhang, L. Zhang, Z. Li, B. Wang, E. Ng, J. Li, and Z. He, "Combining convolutional neural network and distance distribution matrix for identification of congestive heart failure," *IEEE Access*, vol. 6, pp. 39734–39744, 2018.

[16] S. L. Oh, E. Y. k. Ng, R. S. Tan, and U. R. Acharya, "Automated diagnosis of arrhythmia using combination of CNN and LSTM techniques with variable length heart beats," *Comput. Biol. Med.*, vol. 102, no. 1, pp. 278–287, Nov. 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0010482518301446

[17] Z. I. Attia, P. A. Friedman, P. A. Noseworthy, F. Lopez-Jimenez, D. J. Ladewig, G. Satam, P. A. Pellikka, T. M. Munger, S. J. Asirvatham, C. G. Scott, R. E. Carter, and S. Kapa, "Age and sex estimation using artificial intelligence from standard 12-lead ECGs," *Circulat., Arrhythmia Electrophysiol.*, vol. 12, no. 9, Sep. 2019, Art. no. e007284.

[18] M. U. Khan, M. Saad, S. Aziz, J. M. Ch, S. Z. H. Naqvi, and M. A. Qasim, "Electrocardiogram based gender classification," in *Proc. Int. Conf. Electr., Commun., Comput. Eng. (ICECCE)*, Jun. 2020, pp. 1–6.

[19] J.-L. Cabra, D. Mendez, and L. C. Trujillo, "Wide machine learning algorithms evaluation applied to ECG authentication and gender recognition," in *Proc. 2nd Int. Conf. Biometric Eng. Appl. (ICBEA)*, 2018, pp. 58–64.

[20] M. Wiggins, A. Saad, B. Litt, and G. Vachtsevanos, "Evolving a Bayesian classifier for ECG-based age classification in medical applications," *Appl. Soft Comput.*, vol. 8, no. 1, pp. 599–608, Jan. 2008.

[21] P. A. Noseworthy, Z. I. Attia, L. C. Brewer, S. N. Hayes, X. Yao, S. Kapa, P. A. Friedman, and F. Lopez-Jimenez, "Assessing and mitigating bias in medical artificial intelligence: The effects of race and ethnicity on a deep learning model for ECG analysis," *Circulat., Arrhythmia Electrophysiol.*, vol. 13, no. 3, Mar. 2020, Art. no. e007988.

[22] F. Liu, C. Liu, L. Zhao, X. Zhang, X. Wu, X. Xu, Y. Liu, C. Ma, S. Wei, Z. He, and J. Li, "An open access database for evaluating the algorithms of electrocardiogram rhythm and morphology abnormality detection," *J. Med. Imag. Health Informat.*, vol. 8, no. 7, pp. 1368–1373, Jul. 2018.

[23] 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," *Sci. Data*, vol. 7, no. 1, p. 154, Dec. 2020.

[24] 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, and M. A. Reyna, "Classification of 12-lead ECGs: The PhysioNet/computing in cardiology challenge 2020," *Physiolog. Meas.*, vol. 41, no. 12, Dec. 2020, Art. no. 124003, doi: 10.1088/1361-6579/abc960.

[25] J. Zheng, J. Zhang, S. Danioko, H. Yao, H. Guo, and C. Rakovski, "A 12-lead electrocardiogram database for arrhythmia research covering more than 10,000 patients," *Sci. Data*, vol. 7, Feb. 2020.

[26] L. Sweeney, "Only you, your doctor and many others may know," *Technol. Sci.*, vol. 2015092903, p. 29, Sep. 2015. [Online]. Available: https://techscience.org/a/2015092903/

[27] L. Sweeney, *Simple Demographics Often Identify People Uniquely*. Pittsburgh, PA, USA: Carnegie Mellon Univ., Data Privacy, 2000. [Online]. Available: http://dataprivacylab.org/projects/identifiability/

[28] B. Hayes, "Uniquely me," *Amer. Sci.*, vol. 102, pp. 106–109, Mar. 2014. [Online]. Available: https://www.americanscientist.org/article/uniquely-me/

[29] National Data Guardian for Health and Care. (2016). *Review of Data Security, Consent and Opt-Outs*. [Online]. Available: https://www.gov.U.K./government/publications/review-of-data-security-consent-and-opt-outs

[30] T. Pilgrim, *De-Identification: The De-Vil is in the De-Tail*. Chennai, India: Mandarin, Nov. 2016.

[31] R. Gellman, "The deidentification dilemma: A legislative and contractual proposal," *Fordham Intellect. Prop. Media Entertain. Law J.*, vol. 21, p. 33, 2010.

[32] B. M. Knoppers, J. R. Harris, I. Budin-Ljøsne, and E. S. Dove, "A human rights approach to an international code of conduct for genomic and clinical data sharing," *Hum. Genet.*, vol. 133, no. 7, pp. 895–903, Jul. 2014.

[33] Office of Civil Rights. (2015). *Guidance Regarding Methods for De-Identification of Protected Health Information*. [Online]. Available: https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html

[34] K. Benitez and B. Malin, "Evaluating re-identification risks with respect to the HIPAA privacy rule," *J. Amer. Med. Inform. Assoc.*, vol. 17, pp. 77–169, Mar./Apr. 2010.

**ARIN GHAZARIAN** (Member, IEEE) received the Ph.D. degree in computational and data sciences from Chapman University, in 2021. He also has two decades of experience in industry working on large scale machine learning, software, and data analytics projects in domains, such as movie and music entertainment. He is a Lecturer of data science and computer science with California State University Long Beach, Chapman University, and continuing education division of University of California Irvine. His research interests include the application of machine learning to ECG analysis, privacy preserving data analysis, deep learning models, and human–computer interaction.

**JIANWEI ZHENG** received his Ph.D. degree in computational and data sciences from Chapman University, in 2021. His research interest includes the application of machine learning to ECG analysis for cardiovascular disease.

**DANIELE STRUPPA** received the laurea degree in mathematics from the University of Milan, Italy, in 1977, and the Ph.D. degree in mathematics from the University of Maryland, College Park, in 1981. He joined as a Provost at Chapman University, in 2006, he is serving currently as the Chapman University's Thirteenth President. He was a recipient of the Bartolozzi Prize from the Italian Mathematical Union, in 1981, the Matsumae Medal from the Matsumae International Foundation of Tokyo, in 1987, the Prestigious Cozzarelli Prize from the National Academy of Sciences for a paper he coauthored, in 2017.

**CYRIL RAKOVSKI** received the Ph.D. degree in biostatistics from Harvard University, in 2006. He was a Postdoctoral Research Fellow at USC, from 2006 to 2008, and then joined Chapman University where he is currently an Associate Professor. His research interests include the genetics analysis, time series data analysis, machine learning, and statistical modeling.