2021 43rd Annual International Conference of the
IEEE Engineering in Medicine & Biology Society (EMBC)
Oct 31 - Nov 4, 2021. Virtual Conference

Increased Risks of Re-identification For Patients Posed by Deep Learning-Based ECG Identification Algorithms

Arin Ghazarian[1], Jianwei Zheng[2*], Hesham El-Askary[3], Huimin Chu[4], Guohua Fu[5] and Cyril Rakovski [6]

***Abstract*— ECGs analysis is an important tool in cardiac diagnosis. ECG data also have the potential to be used as a biometric source that allows precise person identification similar to the widely used fingerprint and iris recognition techniques. However, this phenomenon also raises serious privacy concerns. In this study, we provide a complete, multi-step ECG identification algorithm using a private database of ECG recordings. We train and validate our AI model on approximately 40k patients which makes this study by far the largest research project in this field. Moreover, our best model attained an exceptionally high accuracy of 94.56%. In addition to discussing the general implications of the deployment of such systems related to privacy, for the first time, we also assess the accuracy of ECG-based identification for distinct heart condition groups (and combinations of such conditions) and the corresponding privacy implications. For instance, we discovered that in contrast to initial expectation that identification accuracy for healthy normal sinus rhythm should be the highest, the identification accuracy is higher for patients with sinus tachycardia or patients who are diagnosed with both ST changes and supraventricular tachycardia. This puts these patients at a higher risk of privacy issues due to re-identification. On the other hand, we observed that patients with premature ventricular contractions have an identification accuracy as low as 78.54%. The identification rate for patients with a pacemaker is 80.2%.**

***Clinical relevance*—While ECG as a biometric can be a potentially useful technology, it also raises serious concerns regarding the privacy of cardiac patients. Especially, the ECG Identification algorithms empowered by deep learning can increase the risk of re-identification.**

Keywords: Arrhythmia, Biometrics, Convolutional neural networks (CNN), Deep Learning, ECG , ECG Identification, Privacy, Re-identification

I. Introduction

Electrocardiogram (ECG) data reflects the bio-electrical activity of the heart collected from human body surface. As shown in Figure 1, an ECG during one normal heartbeat consists of several features including the P-wave, the QRS complex, the T-wave, PR interval, QT interval, PR segment and ST segment. ECG is an important and non-invasive tool in the diagnosis of heart status and detection of abnormalities. The amplitudes, time intervals, and other morphological features in different sections of the ECG signal are used for diagnoses and classification of different types of cardiac conditions. Arrhythmia is a group of conditions in which the heartbeat has an irregular rate or rhythm. Arrhythmia has a wide and significant impact on public health, quality of life, and medical expenditures [1]. For example, the most common type of arrhythmia, atrial fibrillation (AFIB) depicted in Figure 2, is associated with a significant increase in the risk of heart failure, cardiac dysfunction and stroke [1].

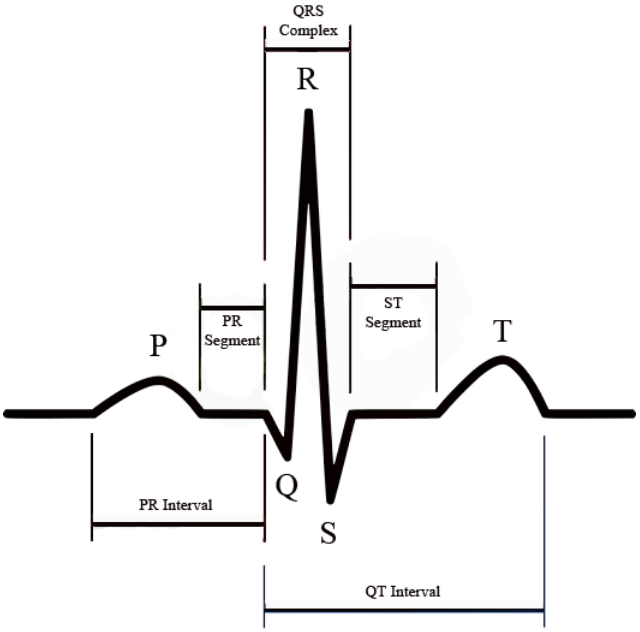

Fig. 1. The ECG waveform and segments in lead II that presents a normal cardiac cycle

ECG data also has the potential to be used as a biometric (electro-physiologic) tool in human identification systems, similar to fingerprint, face, and iris [2], [3]. It also eliminates the aliveness test required in some other forms of biometric since heart signal is an inherently alive biometric. Even though the application of ECG in biometrics is a useful technology, it also raises serious privacy concerns about re-identification of patients via AI-based ECG matching techniques. For example, an ECG-based biometrics system can also diagnose and store heart conditions of the users in the background. Vice versa, the ECG recordings collected

[1] Schmid College of Science and Technology, Chapman University,Orange, CA 92866, USA ghazarian@chapman.edu
[2*]Corresponding author: Jianwei Zheng, Schmid College of Science and Technology, Chapman University, Orange, CA 92866, USA zheng120@mail.chapman.edu
[3]Schmid College of Science and Technology, Chapman University, Orange, CA 92866, USA elaskary@chapman.edu
[4] Ningbo First Hospital, Zhejiang University, Ningbo, China mark.chuhuimin@gmail.com
[5] Ningbo First Hospital, Zhejiang University, Ningbo, China eagle1002@126.com
[6] Schmid College of Science and Technology, Chapman University, Orange, CA 92866, USA rakovski@chapman.edu

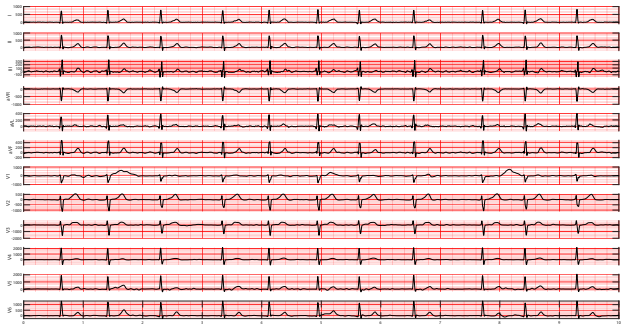

Fig. 2. A 12-lead ECG showing atrial fibrillation rhythm that has no visible P waves that are replaced by coarse fibrillatory waves and an irregularly irregular QRS complex.

from patients for diagnosis or research purposes can be matched with external ECG databases using ECG biometric identification systems to recognize the individual patient. Additionally, privacy becomes even more important as we collect and store a constantly increasing volume of data from the citizens. For example, the emerging medical wearable device technologies capture and store a continuous stream of sensitive health data. The accumulation of such large amounts of personal information makes the privacy protection an even more critical requirement.

II. Literature Review

Automated analysis of ECG data using machine learning techniques has been the focus of many recent cardiac research efforts such as arrhythmia classification [4] and accurate prediction of ventricular arrhythmia origins [5]. Furthermore, ECG data have also been used for emotion recognition [6], [7]. A typical ECG machine learning pipeline includes denoising and baseline correction, heartbeat segmentation and QRS detection, feature extraction, and model training. Both temporal/morphological features like amplitude, duration, or slopes of different sections of the PQRST segment and frequency domain features like Fourier or wavelet transformation coefficients have been used by researchers. For instance, QRS duration and amplitude of the P-wave are features from the time domain and Daubechies discrete wavelet transformation coefficients, wavelet scaleograms, and Fast Fourier Transformation (FFT) coefficients are features from the frequency domain. Support Vector Machines (SVM), naive Bayes, random forest, neural networks and their variations have been some of the commonly used machine learning techniques in ECG research.

The idea of using ECG as a biometric identifier has been around for a long time [8], [2]. Even though, subsequent studies have reported high identification accuracies, all of them were based on small number of subjects, ranging from ten to a few hundred. These studies considered ECG recordings with single and multiple leads. Belo et al [9] leveraged Temporal Convolutional Neural Network (TCNN) and Recurrent Neural Network (RNN) for both ECG identification and authentication. The authors report that overall, the TCNN outperforms the RNN achieving 100%, 96% and 90% accuracy on Fantasia (40 subjects), MIT-BIH (47 subject), and CYBHi (63 subjects) databases respectively. Labati et al [10] used a CNN-based deep learning approach to extract features from ECG and achieved 100% accuracy on around 50 human subjects. Deshmane and Madhe [11] proposed a CNN based approach achieving 81.33%, 96.95%, 94.73%, and 92.85% accuracies on MITDB (47 subject), FANTASIA (40 subjects), NSRDB (18 subjects), and QT databases (105 subjects). Eduardo et al [12] used autoencoders for denoising and feature extraction in an ECG biometric system. Salloum and Kuo [13] used Recurrent Neural Networks (RNN) with Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRU). They could reach 100% identification rate with 90 subjects of the public ECG-ID database. Zhang et al [14], achieved an average identification rate of 93.5% using a multi-resolution CNN on different datasets ranging from 18 to 47 subjects. Li et al [15] used two cascaded CNNs, the first CNN is used for feature extraction from ECG heartbeats and the second one is used for identification. They used a mixed dataset from 184 subjects and achieved 99.52% accuracy.

Table I summarizes the results from previous research in ECG Identification.

The deep leaning model employed to re-identify subjects based on their ECG recordings needs to possess exceptional inherent quality, especially as the number of patients in the data increases. 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 $p*100\%$ 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,

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

Lastly, using Stirling's approximation equation (2) yields,

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

In our study, with sample of size n was 39,754 and attained accuracy of 94.56%, the proportion of such favorable re-identification selection is practically zero.

III. Data

The 12-lead ECG data analyzed in this work came from two open access research databases [16], [17], containing 10,646 and 344 ECG recordings respectively, as well as an additional new dataset from the Ningbo First Hospital, including 34,320 ECG recordings. The institutional review board of Ningbo First Hospital approved this study and granted the waiver of the requirement to obtain informed consent. There are 88 cardiac conditions present in the combined data that contains 45,310 ECG recordings consisting

of 10-second, 12-lead ECGs with 500 Hz sampling rate. Cardiologist-supervised physicians interpreted each recording and gave cardiac condition labels and ECG findings. The number of volts per A/D bit was 4.88, and the A/D converter had 32-bit resolution with upper and lower limits of 32,767 and -32,768 microvolts respectively. Detailed description of the enrolled participants' baseline characteristics and condition frequency distribution can found in our previous work [4].

IV. Preprocessing

We employed a three-stage noise reduction method that includes Butterworth low-pass filter to remove high-frequency noise (above 50 Hz), the Robust LOESS to eliminate baseline wandering and Non-Local Means (NLM) to remove residual noise [4]. The major known sources of noise contamination were power line interference, electrode contact noise, motion artifacts, skeletal muscle contraction, baseline wandering, and random noise. The baseline wandering, the low frequency noise component (<0.5Hz), could be induced by respiration. The high frequency (50-60Hz) noise component majorly was caused by the power line interference. The ECG recordings from patients were broken down into R-peak to R-peak intervals to be used as the input unit to the neural network. Instead of heartbeats we used R-to-R interval since it is easier and more accurate than trying to find the boundary of each heartbeat.

V. Deep Learning Model

As shown in Figure 3, we implemented a Convolutional Neural Network with three repeated sequences of convolution, batch normalization, and max-pooling, followed by a flattening layer, two dense layers, and a final output softmax layer. The number of units in the final softmax layer is equal to the number of patients (38,378) in the dataset. We used relu activation function for all of the convolutional layers. The input vector was a vector of length 12*300 representing the 12 leads data for a single R-to-R interval. The labels were the encoded numbers for the patient IDs in the database. Adam optimizer was used with sparse categorical crossentropy loss function. We trained the model on the GPU machines provided by the Keck computational research cluster at Chapman University

The data consisted of ECG samples from 38,378 patients in both the training and test sets. We randomly selected 20% of the R-to-R intervals from each patient to be used in the validation set and the rest were used in the training set. Thus, all individuals were present in both the training and validation sets but with distinct and non-overlapping R-to-R interval data. There was a total of 497,911 R-to-R intervals, of which 398,328 were used in the training and 99,583 in the test sets respectively (20% of the data was used for validation). We achieved an accuracy of 94.56% (the percent of the total number of R-to-R intervals in the validation data identified correctly). Figure 4 shows the convergence curve from the model training. It is clear that the training and validation curves have a good fit.

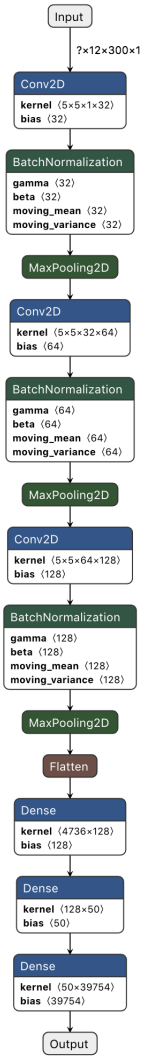

Fig. 3. Deep Learning Model Architecture

VI. Discussion

We designed and implemented an AI algorithm aimed at identifying subjects based on 12-lead ECG data. We extended all existing studies in terms of sample size and were able to attain a high accuracy. As shown in Table I, Previous results had severe limitations due to sample sizes. In this study we showed analytically the exploding complexity of the identification process as the sample size grows. In this study we trained a model with 38,378 subjects in both the training and validation datasets. The accuracy attained by the algorithm was exceptionally high at 94.56%.

A. Privacy Risks

As we saw in the previous sections, advanced AI techniques like CNNs enable us to identify individuals in a large population using their ECG signals. Despite its useful applications, this also brings new privacy and ethical concerns regarding ECG data. For instance, ECG identification now creates the potential for re-identification of individuals in ECG databases. Re-identification is the practice of

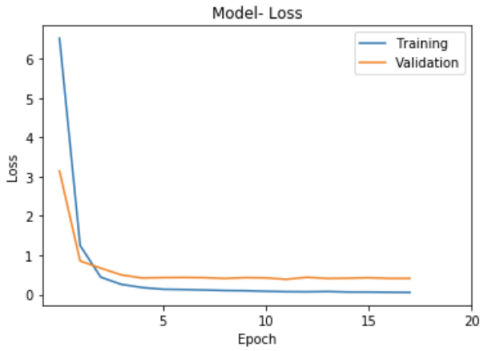

Fig. 4. Training loss convergence curve

discovering the individuals in an anonymized database by matching the records with publicly available information (auxiliary data). Having the wrong belief that anonymizing data protects the privacy of patients; many institutions release anonymized data for research purposes. For example, in 1997, researchers from MIT were able to re-identify the governor of Massachusetts in an anonymized healthcare insurance database released for research purposes [18]. They were able to re-identify him by matching this database with the publicly available voter registration data.

One potential re-identification scenario in ECG datasets can happen when individuals contribute data to two different research databases. For example, database A has ECG sample, gender, and date of birth and database B has an ECG sample and zip code. By simply matching the ECG columns in both databases using an identification system like the one we proposed in this paper, one can discover the individuals which appear in both datasets and obtain a complete profile of the individuals by joining their records. In this case, we have gender and date of birth from database A and zip code from database B. These three demographics attributes might be enough to uniquely identify someone since 87% of US citizens can be uniquely identified only by having their date of birth, gender and zip code [19], [20].

B. Privacy Risks per Condition

To assess the privacy risks for cardiology patients posed by ECG identification technology, we calculated the misidentifications proportions per disease in our validation set. We considered two scenarios in our calculations. In the first scenario, multiple diagnoses per patient were flattened, meaning that if a patient who was misidentified had two conditions, we counted that as misidentification in both category of conditions. In the second scenario, we assumed multiple diagnoses for a patient as a single category to understand the joint effect of multiple conditions on the misidentifcation rate. The results for conditions where we had more than 500 samples are shown in Table II and Table III. Patients with conditions which have high identification accuracy rates should be more concerned with the protection of their ECG data.

From Table II, we can observe that healthy sinus rhythm and conditions like sinus tachycardia, and supraventricular tachycardia have high identification rates, while conditions such as premature ventricular contractions, aberrant ventricular conduction, or patients with pacemaker have low identification rates. The identification rate for patients with a pacemaker is 80.2%. One interesting observation is that in contrast to the common expectation, some conditions have even a better identification rate than normal sinus rhythm. For instance, the highest identification rate was attained for patients diagnosed with both ST changes and supraventricular tachycardia (99.25%). We can also see in Table III that patients who are diagnosed with atrial fibrillation only or a combination of atrial fibrillation and other conditions had a lower identification rate and are at the bottom of the table. On the other hand, we also notice that the top five identification rates at Table III are for the group of patients who have a form of tachycardia. Tachycardia is a group of heart conditions in which heart beat rate is too fast.

C. Future Work

As we shown in the paper, aggregation and anonymization do not guarantee privacy and individuals can be re-identified even from the published aggregated results specially with the ECG data which can uniquely identify people. We are working on the application of privacy preserving techniques like differential privacy to ECG datasets. Differential privacy enables us to share aggregated statistics from private datasets, while preserving individual's privacy [21]. It works by randomly adding noise to the results to protect the privacy of individuals while sacrificing some accuracy in the published analysis.

VII. Acknowledgments

We are grateful for the support of Shaoxing People's Hospital (Shaoxing Hospital Zhejiang University School of Medicine) ECG department.
We are grateful for the medical device support from Zhejiang Cachet Jetboom Medical Devices CO.LTD.
We received the software engineering support provided by Kelvin Zheng and Terence Wang from Global Customer Support of Schneider Electric Software.

TABLE I: Results from Previous Research

Research	**Number of Subjects**	**Accuracy**
Belo et al [9]	40, 47, 63 subjects)	100%, 96% and 90%
Deshmane and Madhe [11]	47, 40, 18, 105	81.33%, 96.95%, 94.73%, and 92.85%
Salloum and Kuo [13]	90	100%
Zhang et al [14]	18-47	93.5%
Li et al [15]	184	99.52%

TABLE II: Identification rate per single condition

Condition Name	**Number of R-to-R Intervals in Validation**	**Identification Rate**
supraventricular tachycardia	3244	98.55
sinus tachycardia	21120	97.14
early repolarization	734	97.0
sinus rhythm	17108	96.98
sinus bradycardia	25573	95.67
counterclockwise vectorcardiographic loop	1371	95.55
tall P wave	847	95.51
ST elevation	2230	95.47
tall tented T wave	906	95.14
1st degree av block	2173	94.85
right axis deviation	2438	94.75
atrial flutter	6377	93.9
st changes	16683	93.8
left ventricular high voltage	11570	93.54
left ventricular hypertrophy	2069	93.23
incomplete right bundle branch block	821	93.06
low QRS voltages	2754	92.85
left anterior fascicular block	1203	92.77
prolonged QT interval	888	92.68
T wave abnormal	14706	92.54
T wave inversion	8194	92.51
ST depression	4571	92.5
sinus arrhythmia	5154	92.39
Q wave abnormal	2572	92.03
complete right bundle branch block	3480	91.87
right bundle branch block	634	91.8
poor R wave progression	2020	91.58
left axis deviation	3675	91.56
Clockwise vectorcardiographic loop	928	91.27
complete left bundle branch block	785	90.32
intraventricular block	1149	89.99
atrial fibrillation	20718	88.92
atrial tachycardia	1022	88.06
premature atrial contraction	3201	87.19
aberrant ventricular conduction	2251	82.14
pacing rhythm	2273	80.2
premature ventricular contractions	3486	78.54

TABLE III: Identification rate considering joint conditions

Condition Name	**Number of R-to-R Intervals in Validation**	**Identification Rate(%)**
ST changes and supraventricular tachycardia	937	99.25
supraventricular tachycardia	900	99.0
sinus tachycardia and T wave abnormal	1187	98.9
sinus tachycardia	8403	98.55
ST changes and sinus tachycardia	1641	97.87
sinus rhythm	12335	97.58
atrial flutter and ST changes	921	97.07
sinus bradycardia	13777	96.66
sinus rhythm and T wave abnormal	680	96.62
left ventricular high voltage and sinus bradycardia	1992	96.44
sinus bradycardia and T wave abnormal	1045	96.27
atrial flutter	979	94.99
sinus arrhythmia	2556	93.62
atrial fibrillation and ST changes	1381	93.12
sinus arrhythmia and sinus bradycardia	653	92.8
atrial fibrillation and left ventricular high voltage	844	91.82
atrial fibrillation and T wave abnormal	1631	91.48
atrial fibrillation	3922	91.46
atrial fibrillation and left ventricular high voltage and ST changes	699	91.13

REFERENCES

[1] S. S. Virani, A. Alonso, E. J. Benjamin, M. S. Bittencourt, C. W. Callaway, A. P. Carson, A. M. Chamberlain, A. R. Chang, S. Cheng, F. N. Delling, L. Djousse, M. S. Elkind, J. F. Ferguson, M. Fornage, S. S. Khan, B. M. Kissela, K. L. Knutson, T. W. Kwan, D. T. Lackland, T. T. Lewis, J. H. Lichtman, C. T. Longenecker, M. S. Loop, P. L. Lutsey, S. S. Martin, K. Matsushita, A. E. Moran, M. E. Mussolino, A. M. Perak, W. D. Rosamond, G. A. Roth, U. K. Sampson, G. M. Satou, E. B. Schroeder, S. H. Shah, C. M. Shay, N. L. Spartano, A. Stokes, D. L. Tirschwell, L. B. VanWagner, and C. W. Tsao, "Heart disease and stroke statistics-2020 update: A report from the american heart association," *Circulation*, vol. 141, no. 9, pp. e139–e596, 2020.

[2] L. Biel, O. Pettersson, L. Philipson, and P. Wide, "Ecg analysis: a new approach in human identification," *IEEE Transactions on Instrumentation and Measurement*, vol. 50, no. 3, pp. 808–812, 2001.

[3] C. Carreiras, A. Lourenço, A. Fred, and R. Ferreira, "Ecg signals for biometric applications - are we there yet?," in *2014 11th International Conference on Informatics in Control, Automation and Robotics (ICINCO)*, vol. 02, pp. 765–772, 2014.

[4] 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, 2020.

[5] 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," *Frontiers in Psychology*, vol. s, 2021.

[6] 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 Transactions on Affective Computing*, vol. 3, no. 1, pp. 18–31, 2012.

[7] S. Brás, J. H. T. Ferreira, S. C. Soares, and A. J. Pinho, "Biometric and emotion identification an ecg compression based method," *Frontiers in Psychology*, vol. 9, p. 467, 2018.

[8] G. Forsen, M. Nelson, R. Staron, P. ANALYSIS, and R. C. R. N. Y., *Personal Attributes Authentication Techniques*. Defense Technical Information Center, 1977.

[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, 2020.

[10] R. Donida Labati, E. Muñoz, V. Piuri, R. Sassi, and F. Scotti, "Deep-ecg: Convolutional neural networks for ecg biometric recognition," *Pattern Recognition Letters*, vol. 126, pp. 78 – 85, 2019. Robustness, Security and Regulation Aspects in Current Biometric Systems.

[11] M. Deshmane and S. Madhe, "Ecg based biometric human identification using convolutional neural network in smart health applications," in *2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA)*, pp. 1–6, 2018.

[12] A. Eduardo, H. Aidos, and A. Fred, "Ecg-based biometrics using a deep autoencoder for feature learning - an empirical study on transferability," in *ICPRAM*, 2017.

[13] R. Salloum and C. . J. Kuo, "Ecg-based biometrics using recurrent neural networks," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2062–2066, 2017.

[14] 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.

[15] 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, 2020.

[16] 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," *Scientific Data*, vol. 7, 02 2020.

[17] J. Zheng, G. Fu, K. Anderson, H. Chu, and C. Rakovski, "A 12-lead ecg database to identify origins of idiopathic ventricular arrhythmia containing 334 patients," *Scientific Data*, vol. 7, no. 1, p. 98, 2020.

[18] L. Sweeney, "Only you, your doctor and many others may know," *Technology Science*, 2015.

[19] L. Sweeney, "Simple demographics often identify people uniquely." 2000.

[20] B. Hayes, "Uniquely me," *American Scientist*, vol. 102, pp. 106–109, 2014.

[21] C. Dwork and A. Roth, *The Algorithmic Foundations of Differential Privacy*. 2014.