Using deep convolutional neural networks to predict patients age based on ECGs from an independent test cohort

Bjørn-Jostein Singstad*[1] and Belal Tavashi[2]

[1]Simula Research Laboratory
[2]Department of Biomedical Engineering, Ankara University

Abstract

Electrocardiography is one of the most frequently used methods to evaluate cardiovascular diseases. However, the last decade has shown that deep convolutional neural networks (CNN) can extract information from the electrocardiogram (ECG) that goes beyond traditional diagnostics, such as predicting a persons age. In this study, we trained two different 1-dimensional CNNs on open datasets to predict age from a persons ECG.

The models were trained and validated using 10 seconds long 12-lead ECG records, resampled to 100Hz. 59355 ECGs were used for training and cross-validation, while 21748 ECGs from a separate cohort were used as the test set. We compared the performance achieved on the cross-validation with the performance on the test set. Furthermore, we used cardiologist annotated cardiovascular conditions to categorize the patients in the test set in order to assess whether some cardiac condition leads to greater discrepancies between CNN-predicted age and chronological age.

The best CNN model, using an Inception Time architecture, showed a significant drop in performance, in terms of mean absolute error (MAE), from cross-validation on the training set (7.90 ± 0.04 years) to the performance on the test set (8.3 years). On the other hand, the mean squared error (MSE) improved from the training set (117.5 ± 2.7 years^2) to the test set (111 years^2). We also observed that the cardiovascular condition that showed the highest deviation between predicted and biological age, in terms of MAE, was the patients with pacing rhythm (10.5 years), while the patients with prolonged QT-interval had the smallest deviation (7.4 years) in terms of MAE.

This work contributes to existing knowledge of age prediction using deep CNNs on ECGs by showing how a trained model performs on a test set from a separate cohort to that used in the training set.

1 Introduction

The electrocardiogram (ECG) was invented by Willem Einthoven in 1901 and since then it has been one of the most important and most frequently used diagnostic tools for cardiovascular diseases. In the 1950s it became possible to convert analog ECG signals to digital ECG signals, this enabled digital interpretation algorithms in the 1960s [1]. These algorithms have generally used rule-based processing techniques to extract features from the ECG in order to classify a large variety of diseases. However, in the last decade, approaches using deep neural networks (DNN) have shown promising performance and present a paradigm shift in how ECGs are being analyzed.

In addition to diagnostic classification, there have been several examples of usage that goes beyond traditional ECG analysis, such as predicting atrial fibrillation in asymptomatic patients [2], risk of death [3], gender and age [4, 5, 6, 7].

In particular, [4] showed that age, predicted by a deep convolutional neural network (CNN), might correlate more with the persons physiological age than the persons biological age. Meaning that, in cases where the predicted age was much higher than the persons biological age, may suggest an underlying disease and might be a biomarker for increased risk of mortality. Furthermore, Lima et al 2021 confirmed, on a separate data set, that the predicted age could be used as a biomarker of the risk of death [5]. However, previous studies have trained and validated the algorithms on ECG from patients admitted to the same hospitals. This approach might overestimate the performance of the model, and to mitigate this the model should be tested on a separate data set from another hospital. In addition, current studies have either just analyzed predictions from small subsets of patients or looked at high-level risk factors when concluding that CNN-predicted age can be used as a biomarker for disease and mortality. This study, therefore, set

*Corresponding Author: b.j.singstad@fys.uio.no

https://doi.org/10.7557/18.6814

out to train a CNN and validate it on ECGs from a completely separate test set. Furthermore, we will compare the predicted age with the biological age for all ECGs in the test set and categorize the ECGs based on cardiologist-annotated cardiovascular diseases.

2 Methods

2.1 Data

12-lead ECG recordings from six different open access data bases [8, 9, 10, 11, 12, 13] was used to train the proposed model in this study. A seventh data set, collected from another hospital, PTB-XL [14] was used as an independent test set. Initially, the training set contained 65900 ECGs and the test set contained 21837 ECGs. After excluding ECGs longer or shorter than 10 seconds and ECGs missing information regarding age or gender the training set contained 59355 ECGs and the test set contained 21748 ECGs. Figure 2 illustrates the exclusion process. The distributions of the patient's age in the training and test set, after the exclusions, are shown in Figure 1.

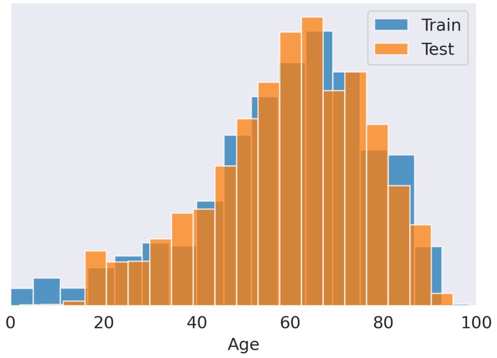


Figure 1: Normalized age distribution of the patients in the training and test set.

In addition to the patient's ECGs, the databases used in this study also contain information about the patient's age, sex as well as cardiologist-annotated cardiovascular conditions. The age was used as the label to predict by the CNNs, and the cardiovascular conditions were used to categorize the predicted age versus the true age on the test set. Table 1 summarizes the cardiovascular conditions considered in this study and the prevalence of each condition in the test set.

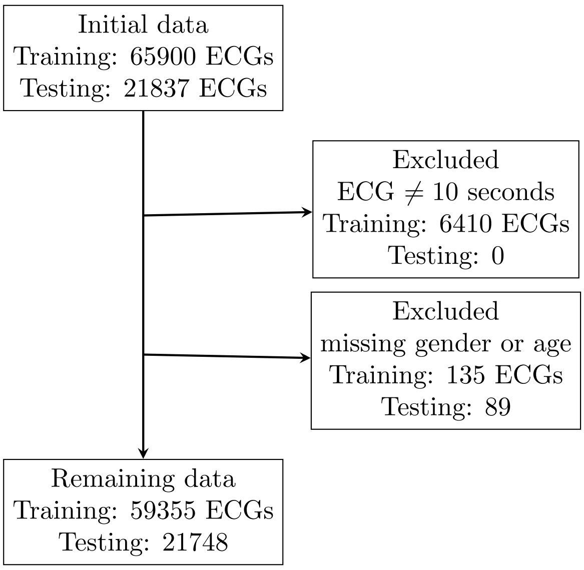


Figure 2: Patients with an ECG recording shorter or longer than 10 seconds or had missing information about gender or age were excluded from the training data.

2.2 Preprocessing

The ECGs were recorded with different electrocardiographs using different sampling frequencies ranging from 256 Hz to 1000 Hz. In this study, we resampled all ECGs to 100 Hz.

2.3 Model

Attia et al 2019 proposed a 1-dimensional CNN to predict age from ECGs[4]. In this study, we compared the Attia model with a model using an Inception Time architecture [15].

2.4 Validation and testing

The models were first evaluated on the training set using 3-fold stratified cross-validation. The stratification was done based on the patient's age and gender. Finally, the models were trained on the entire training set and then applied to the test set.

Diagnoses	Prevalence	Diagnoses	Prevalence
Pacing Rhythm	390	Premature Atrial Contraction	396
Prolonged QT Interval	119	Left Axis Deviation	5216
Atrial Fibrillation	1682	Sinus Bradycardia	696
Atrial Flutter	88	Sinus Rhythm	19640
Left Bundle Branch Block	542	Sinus Tachycardia	1073
Q Wave Abnormal	540	Sinus Arrhythmia	1004
T Wave Abnormal	2427	Left Anterior Fascicular Block	2037
Prolonged PR Interval	329	Right Axis Deviation	416
Low QRS Voltage	185	T Wave Inversion	385
1st Degree AV block	801	Supraventricular Premature Beats	209

Table 1: The prevalence of the 20 cardiologist annotated cardiovascular conditions in the test set after exclusions.

The models were trained using Google Colab with 12 GB GPU and a CPU with 25GB RAM.

2.5 Hyperparameters

The initial hyperparameters were set equal to the best hyperparameters found in [16]. Furthermore, we evaluated the performance during development by testing batch sizes in the range from 16 to 64 and found 16 to be optimal. As in [16] we used 0.001 as the initial learning rate and in addition, we designed a learning rate scheduler using automatic learning rate reduction during development which reduced the learning rate with a factor of 10 each time. We found the automatic learning rate reduction to happen most frequently at the 10th and 15th epoch. For the final training before applying the model on the test set we then developed a learning rate scheduler with a fixed learning rate reduction at the 10th and the 15th epoch. To determine the total number of epochs to use for the final training we used the training curves from the development. 20 epochs were selected for the final training because we observed, from the training curves during development shown in Figure 4, that there was no decrease in loss on the validation data after 20 epochs.

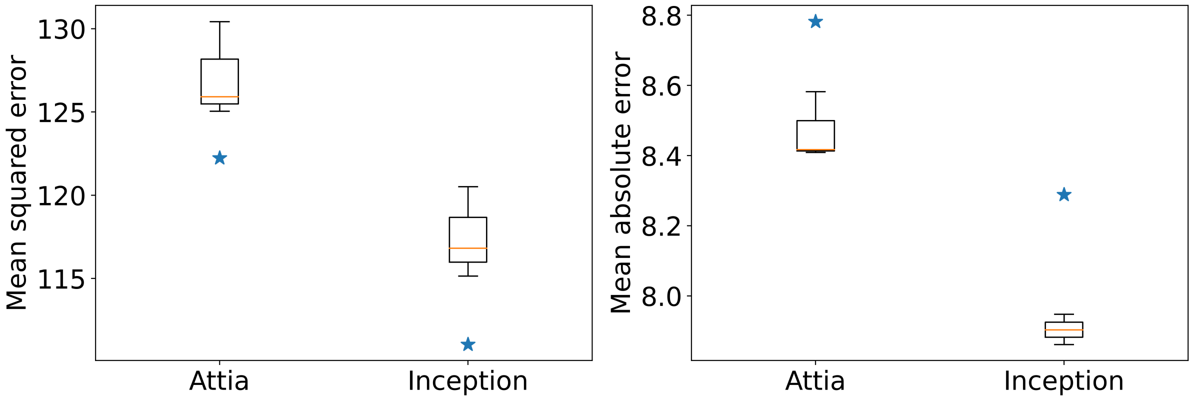


Figure 3: The box and whisker plots represent the mean squared error and the mean absolute error achieved by the Attia model and the Inception Time model using 3-fold cross-validation on the training set. The stars represent the scores obtained on the test set.

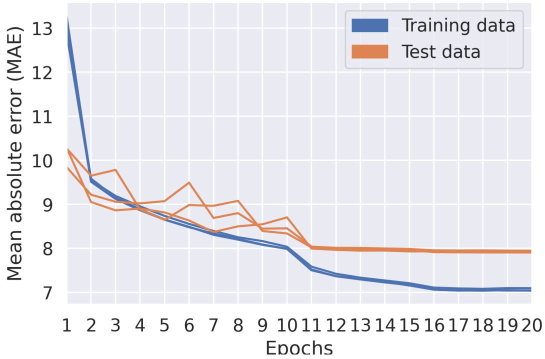


Figure 4: Loss curves showing the models performance on the training and validation folds during 3-fold cross-validation on the training set

3 Results

To compare the difference between the Attia model and the Inception Time model we present the difference, in terms of mean absolute error (MAE) and mean squared error (MSE), on the training and test set in Figure 3. The results achieved, using 3-fold cross-validation on the training set, are represented as box and whisker plots, while the performance on the test set is represented with a single point (a star). The Attia model achieved a cross-validated MAE of 8.46 ± 0.09 years and a MSE of 127 ± 2.9 year^2 on the training set and a MAE of 8.78 years and a MSE of 122.2 year^2. The Inception model achieved a cross-validated MAE of 7.90 ± 0.04 years and a MSE of 117.5 ± 2.7 year^2 on the training set and a MAE of 8.3 years and a MSE of 111 year^2.

Figure 5 provides the relationship between the biological and DNN-predicted age for all of the 21748 patient ECGs in the test set. True versus predicted age by the Attia model are shown in Figure 5a and Inception Time model in Figure 5b. The red line in both figures represents the optimal age prediction, while the green line shows the optimal linear fit between predicted age and true biological age using linear regression.

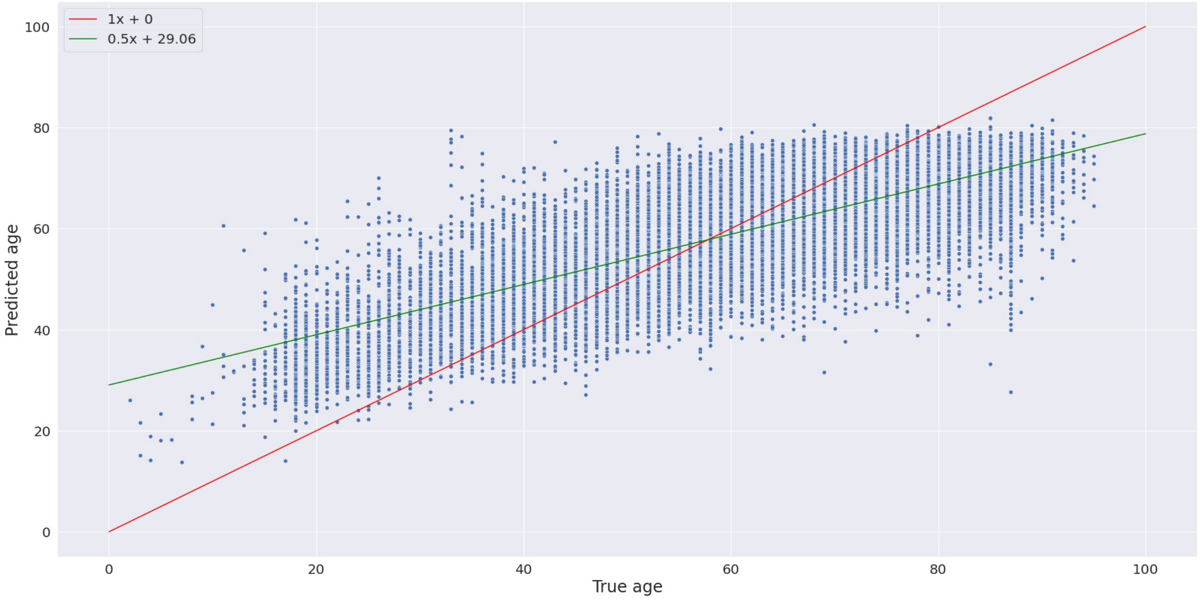
Figure 6 show the CNN predicted age versus the true biological age on the test set, categorized based on the 20 cardiologist-annotated cardiovascular conditions. As in Figure 5 the red lines represent the optimal age prediction, while the green lines show the optimal linear fit between predicted age and true biological age using linear regression. In addition, the MAE for each category is given in the header of each subplot in Figure 6.

4 Discussion

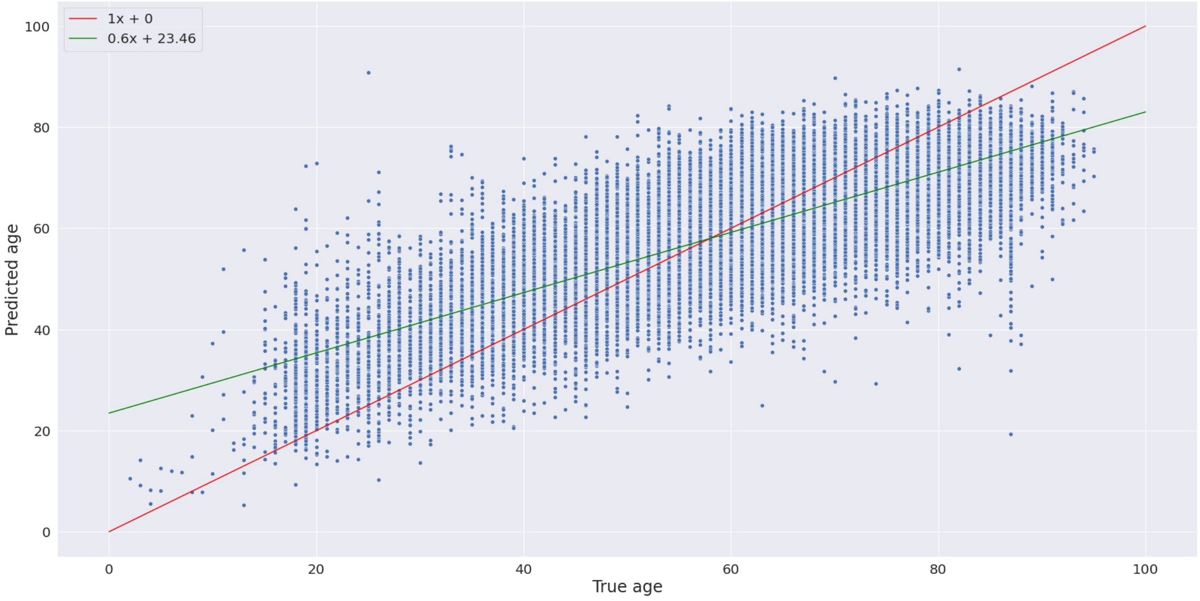
The findings in this study broadly support the work of other studies in this area, linking ECG with age prediction. In the current study, we trained two CNNs, one proposed by Attia et al 2019 with a second model called Inception Time, to predict age from a persons ECG. Furthermore, we compared the predicted age with the true biological age across all patients in the test set and we also compared them categorized based on cardiologist-annotated cardiovascular diagnoses.

Figure 3 shows that the Inception Time model performed significantly better than the Attia model both on the training and the test set. However, from Figure 3 we also see that both models had a significant drop in performance in terms of MAE from training to test set, but it is somewhat surprising that both models also improved the MSE significantly on the test set compared to the training set. The drop in MAE was probably caused by the fact that the models were tested on a data set with patients admitted to a completely different hospital and with a slightly different age distribution, as seen from Figure 1. To understand the improvement in performance in terms of MSE from training to the test set we have to keep in mind that MSE first and foremost punishes large prediction errors. Thus, a possible explanation might be understood by looking at the comparison of true and predicted age in Figure 5 and the age distribution in the train and test set shown in Figure 1. From Figure 5 we see that the biggest discrepancies between true and predicted age are located at each end of the age scale, which is consistent with the reported results in [4], and by looking at Figure 1 we see that, in contrast to the training set, there are almost no patients < 20 years in the test set, and this might reduce the number of prominent outliers in the test set predictions.

From a clinical point of view, large outliers and high MSE on the test set could be caused by having patients with more serious CVDs or healthier persons than what was present in the training set. Therefore, evaluating such a model based on MSE



(a) Attia



(b) Inception Time

Figure 5: The figures show the relationship between the deep neural network-predicted age and the true (biological) age. The red line shows the best fit for an optimal model, while the green line shows the best linear fit for the current DNN models.

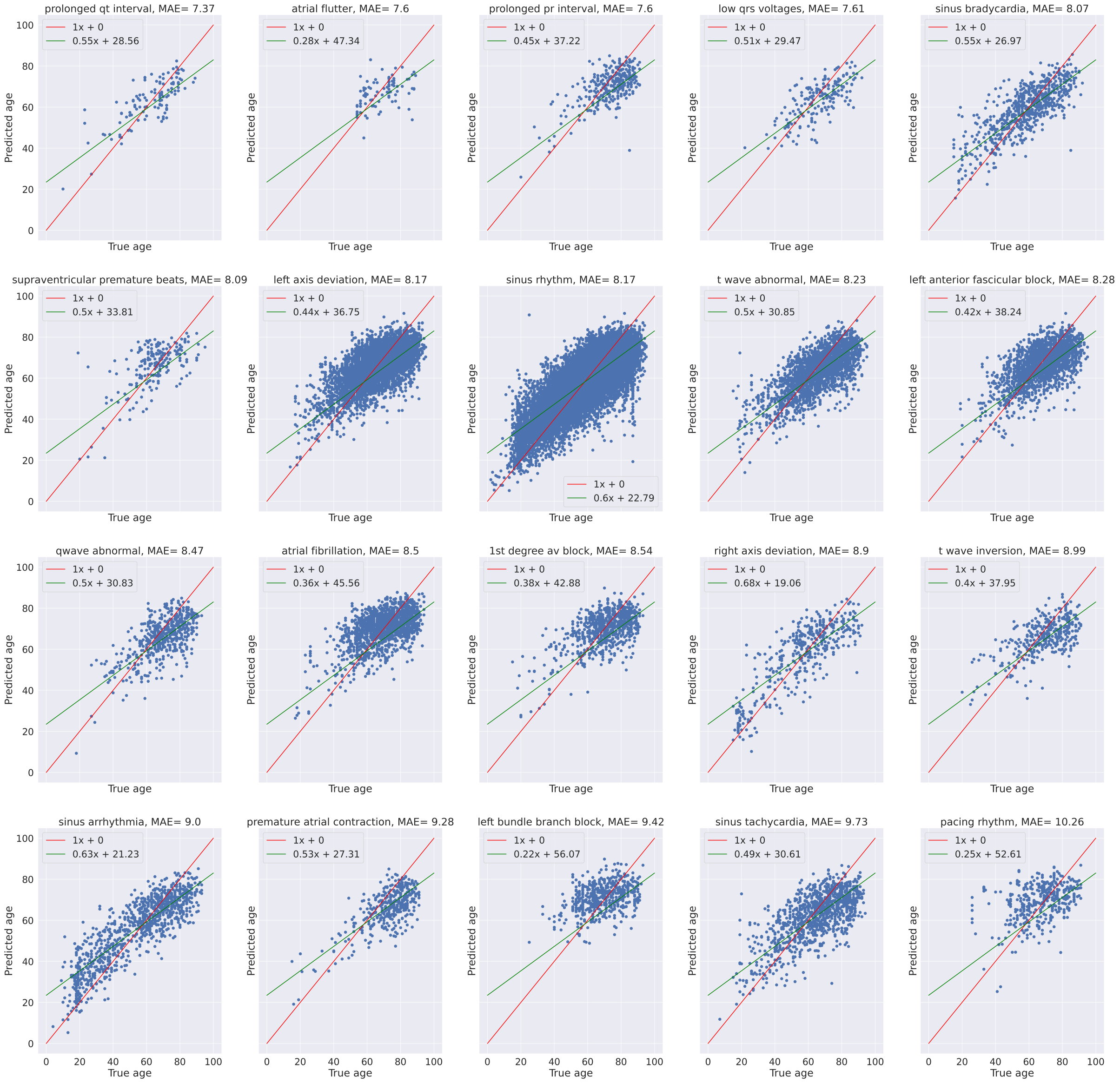


Figure 6: Predicted age from the Inception Time model versus true (biological) age of the patient in the test set. The comparison of predicted and true age are categorized into 20 groups of different cardiovascular conditions. The groups are sorted, from the upper left corner to the lower right corner, based on the mean absolute error (MAE) between the predicted and the true age.

should be done with caution. MAE on the other hand will weigh each prediction error equally and would therefore be less affected by a few outliers.

In Figure 6 we have categorized the patients based on cardiologist-annotated cardiovascular conditions, and within each category, we compare the true and predicted age. Contrary to our expectations, the group with the highest MAE (prolonged QT-interval (LQT) MAE= 7.36) was not so different from the group with the lowest MAE (Pacing rhythm MAE= 10.47). In addition, we hypothesized that there would be a linear relationship between the severity of the diagnosis and the degree of misinterpretation in terms of MAE. However, LQT and left bundle branch block (LBBB) are located on each side of the MAE scale, both associated with an increased risk of mortality in contrast to sinus rhythm, for instance, which is considered to be the healthy class in these datasets, but has a MAE less than LBBB and greater than LQT.

A limitation of this study is that the datasets used, both in training and testing, only contain ECGs from patients admitted to the hospital. Even though the ECGs in the training and test set are recorded from different hospitals, from different countries and in some cases with different electrocardiographs, it should still be kept in mind that a large portion of the patients has some sort of disease, since they are admitted to the hospital. In future work, it would be interesting to train a model on a hospital cohort and test it on an independent healthy cohort or vice versa to see how this affects the performance.

5 Conclusion

The main goal of the current study was to investigate the performance of a CNN-based age predictor, when tested on a test set from a separate cohort, and compare it to the performance using CV on the training set. The results of this investigation showed that both CNN models tested had a relatively small but significant drop in performance in terms of MAE. This study, therefore, confirms the findings by [4], who showed that CNNs could be used to predict a persons age, but also emphasizes the importance of testing such models on separate test sets in order to keep control of possible biases acquired by the model and we therefore strongly suggest that future studies in this field report results on independent test sets in addition to the performance on the training set using some kind of resampling method.

6 Code Availability

The code that was used to implement the model and produce the results presented in this paper is hosted on GitHub: https://github.com/Bsingstad/ECG-age

References

[1] Harold Smulyan. The Computerized ECG: Friend and Foe. *The American Journal of Medicine*, 132(2):153–160, February 2019. doi: 10.1016/j.amjmed.2018.08.025.

[2] Zachi I Attia et al. An Artificial Intelligence-Enabled ECG Algorithm for the Identification of Patients with Atrial Fibrillation during Sinus Rhythm: a Retrospective Analysis of Outcome Prediction. *The Lancet*, 394 (10201):861–867, September 2019. doi: 10.1016/S0140-6736(19)31721-0.

[3] Sushravya Raghunath. Prediction of mortality from 12-lead electrocardiogram voltage data using a deep neural network. *Nature Medicine*, 26(6):886–891, June 2020. ISSN 1546-170X. doi: 10.1038/s41591-020-0870-z. URL `https://www.nature.com/articles/s41591-020-0870-z`. Number: 6 Publisher: Nature Publishing Group.

[4] Zachi I. Attia et al. Age and Sex Estimation Using Artificial Intelligence From Standard 12-Lead ECGs. *Circulation: Arrhythmia and Electrophysiology*, 12(9):e007284, September 2019. doi: 10.1161/CIRCEP.119.007284. Publisher: American Heart Association.

[5] Emilly M. Lima et al. Deep neural network-estimated electrocardiographic age as a mortality predictor. *Nature Communications*, 12(1):5117, August 2021. ISSN 2041-1723. doi: 10.1038/s41467-021-25351-7. URL `https://www.nature.com/articles/`

s41467-021-25351-7. Number: 1 Publisher: Nature Publishing Group.

[6] Chiao-Hsiang Chang, Chin-Sheng Lin, Yu-Sheng Luo, Yung-Tsai Lee, and Chin Lin. Electrocardiogram-Based Heart Age Estimation by a Deep Learning Model Provides More Information on the Incidence of Cardiovascular Disorders. *Frontiers in Cardiovascular Medicine*, 9, 2022. ISSN 2297-055X. URL https://www.frontiersin.org/articles/10.3389/fcvm.2022.754909.

[7] Adetola O Ladejobi et al. The 12-lead electrocardiogram as a biomarker of biological age. *European Heart Journal - Digital Health*, 2(3): 379–389, September 2021. ISSN 2634-3916. doi: 10.1093/ehjdh/ztab043. URL https://doi.org/10.1093/ehjdh/ztab043.

[8] Matthew A Reyna, Nadi Sadr, Erick A Perez Alday, Annie Gu, Amit J Shah, Chad Robichaux, Ali Bahrami Rad, Andoni Elola, Salman Seyedi, Sardar Ansari, Hamid Ghanbari, Qiao Li, Ashish Sharma, and Gari D Clifford. Will Two Do? Varying Dimensions in Electrocardiography: The PhysioNet/Computing in Cardiology Challenge 2021. In *2021 Computing in Cardiology (CinC)*, volume 48, pages 1–4, September 2021. doi: 10.23919/CinC53138.2021.9662687. ISSN: 2325-887X.

[9] Erick A. Perez Alday, Annie Gu, Amit J. Shah, Chad Robichaux, An-Kwok Ian Wong, Chengyu Liu, Feifei Liu, Ali Bahrami Rad, Andoni Elola, Salman Seyedi, Qiao Li, Ashish Sharma, Gari D. Clifford, and Matthew A. Reyna. Classification of 12-lead ECGs: the PhysioNet/Computing in Cardiology Challenge 2020. *Physiological Measurement*, 41(12):124003, December 2020. ISSN 0967-3334. doi: 10.1088/1361-6579/abc960. URL https://dx.doi.org/10.1088/1361-6579/abc960. Publisher: IOP Publishing.

[10] Feifei Liu et al. An Open Access Database for Evaluating the Algorithms of Electrocardiogram Rhythm and Morphology Abnormality Detection. *Journal of Medical Imaging and Health Informatics*, 8(7):1368–1373, September 2018. doi: 10.1166/jmihi.2018.2442.

[11] Vikto Tihonenko, A Khaustov, S Ivanov, A Rivin, and E Yakushenko. St Petersburg INCART 12-lead arrhythmia database. *PhysioBank PhysioToolkit and PhysioNet*, 2008.

[12] R. Bousseljot, D. Kreiseler, and A. Schnabel. Nutzung der EKG-Signaldatenbank Cardiodat der PTB über das Internet. *Biomedizinische Technik/Biomedical Engineering*, pages 317–318, July 2009. doi: 10.1515/bmte.1995.40.s1.317.

[13] Jianwei Zheng and et al. Optimal Multi-Stage Arrhythmia Classification Approach. *Scientific Reports*, 10(1):2898, February 2020. doi: 10.1038/s41598-020-59821-7.

[14] Patrick Wagner et al. PTB-XL, a Large Publicly Available Electrocardiography Dataset. *Scientific Data*, 7(1):154, May 2020. doi: https://doi.org/10.1038/s41597-020-0495-6.

[15] Hassan Ismail Fawaz et al. InceptionTime: Finding AlexNet for Time Series Classification. *Data Mining and Knowledge Discovery*, 34(6):1936–1962, November 2020. ISSN 1573-756X. doi: https://doi.org/10.1007/s10618-020-00710-y. URL https://doi.org/10.1007/s10618-020-00710-y.

[16] Bjørn-Jostein Singstad and Eraraya Morenzo Muten. Assessing the Impact of Downsampled ECGs and Alternative Loss Functions in Multi-Label Classification of 12-Lead ECGs. November 2022. doi: 10.1101/2022.11.16.22282373. URL http://medrxiv.org/lookup/doi/10.1101/2022.11.16.22282373.