Predicting Age and Sex from Reduced Lead Electrocardiograms using Deep Learning

Felipe M. Dias[1,2], **Estela Ribeiro**[1], **Quenaz B. Soares**[1], **Jose E. Krieger**[1], **Marco A. Gutierrez**[1,2]

[1]Heart Institute (InCor) – Clinics Hospital
University of Sao Paulo Medical School (HCFMUSP)
Sao Paulo – SP – Brazil

[2]Polytechnique School, University of Sao Paulo (POLI USP)
Sao Paulo – SP – Brazil

f.dias@hc.fm.usp.br, estela.ribeiro@hc.fm.usp.br, quenaz.soares@hc.fm.usp.br

j.krieger@hc.fm.usp.br, marco.gutierrez@incor.usp.br

Abstract. *Artificial intelligence is increasingly used to extract health insights from 12-lead (12L) electrocardiograms (ECG). Here, we propose a deep-learning model to predict sex and age from 12L and reduced-lead ECGs (6L–1L) and assess their impact on mortality risk. Using a ResNeXt-based model trained on the CODE15 dataset, our best models achieved an F1-score of 0.800 for sex classification (12L) and a mean absolute error of 8.961 for age estimation (4L). We found that overestimated age predictions and incorrect sex classifications were associated with higher mortality risk, whereas underestimated age predictions correlated with lower risk. These findings highlight the potential of reduced-lead ECGs for risk assessment, expanding their clinical utility.*

1. Introduction

The electrocardiogram (ECG) is an important diagnostic tool used in modern medicine, enabling the diagnosis of several heart conditions, including myocardial infarction and atrial fibrillation [AlGhatrif and Lindsay 2012]. It is also widely accessible, non-invasive, and relatively inexpensive. Due to its ease of use and high diagnostic value, the ECG has become an integral part of routine medical practice and is used in a wide range of clinical settings, from primary care to emergency departments and intensive care units [Rafie et al. 2021].

The use of Artificial Intelligence (AI) for automatic analyzing ECG signals dates back over 60 years, with notable works [PIPBERGER et al. 1960]. These studies paved the way for the development of automatic ECG classification methods [Macfarlane and Kennedy 2021]. Automatic ECG classification systems are usually designed to perform physician tasks, such as disease diagnosis [Cohen-Shelly et al. 2021], but AI has recently demonstrated capabilities beyond human expertise [Attia et al. 2019b].

However, these advancements often rely on 12-lead ECGs, restricting their use to clinical settings. However, the growing availability of reduced-lead devices, such as the Apple Watch (1-lead) and AliveCor (1-lead, 6-lead), presents new opportunities for AI-driven diagnostics, expanding accessibility to arrhythmia detection and other applications.

Age and sex factors are known to influence ECG [Macfarlane et al. 1994]. The patterns presented on ECGs are different among individuals [Batchvarov et al. 2002], and are sex [Malik et al. 2013] and age dependent [Macfarlane et al. 1994]. In this context, recent works propose that age and sex can be estimated by 12-lead ECG exams. For instance, [Attia et al. 2019a] used a private dataset to predict age and sex from 12-lead ECG signals with a DL model. Sex estimation was more accurate in younger subjects. Age estimation indicates that there is a correlation between the gap of age predicted by the model and the actual chronological age with the incidence of cardiovascular diseases. [Lima et al. 2021] used an extensive 12-lead ECG dataset to predict age with a ResNet DL model, proposing that their ECG-age estimator can be seen as a predictor of mortality.

Our work proposes an approach for predicting age and sex from reduced lead ECGs. We investigated the performance of several ECG lead-set configurations (1L, 2L, 3L, 4L, 6L, and 12L) in both tasks. We employed a network based on the ResNeXt [Xie et al. 2017b] architecture and trained our approach for 30 epochs using the Adam optimizer. We used the CODE15 dataset, the largest publicly available ECG dataset, for training and evaluation. To ensure the reliability of our results, we performed a 70%/30% train-test split and used a 5-fold cross-validation approach on the training set. Our reported results are the mean and standard deviation of the 5-fold trained networks on the test set. Also, we evaluated if there is an increased risk of mortality among individuals that our trained model for age estimation differs by a certain threshold. We make a similar analysis using our trained model for sex identification and evaluate if there is an increase in mortality among individuals where our model misclassifies the subjects' sex.

The contributions of this study can be listed as follows:

1. Development of a 1D-based ResNeXt architecture trained for age and sex prediction using electrocardiogram signals.
2. Application of age and sex prediction methods on reduced lead ECG configurations (1L, 2L, 3L, 4L, 6L, and 12L), demonstrating their potential usability with devices that utilize reduced lead ECGs, such as Apple Watch and AliveCor.
3. Demonstration that our age and sex prediction models can serve as effective predictors of mortality for 12L ECGs and reduced lead ECGs.

2. Material and Methods

2.1. Dataset

We used the dataset from the Clinical Outcomes in Digital Electrocardiography (CODE) study, obtained from the TeleHealth Network of Minas Gerais (TNMG) in Brazil [Ribeiro et al. 2020]. The dataset consists of digital 12-lead ECG exams collected between 2010 and 2016. These exams were obtained from 811 cities in the state of Minas Gerais, Brazil. In total, the dataset includes clinical data and 12-lead ECG recordings from 1,558,415 patients. The ECG signals in the dataset were sampled at a frequency of 400 Hz, with a duration of either 7 or 10 seconds. To standardize the signal length, zero-padding was applied to both the beginning and end of the original signal. This padding continued until the signals reached a size of 4,096.

In this work, we used the CODE15 dataset which was created by selecting 15% of the patients from the CODE study, resulting in 345,779 exams from 233,770 patients

[Lima et al. 2021]. Unlike the CODE study, which is not publicly available, this dataset can be accessed and used by anyone as it is licensed under the Creative Commons Attribution 4.0 International license (CC BY 4.0.). The dataset includes metadata for each exam, including the exam IDs, patient ID, age, and sex. Additionally, the metadata includes information on whether the patient has 1st-degree AV block, right or left bundle branch block, sinus bradycardia, atrial fibrillation, sinus tachycardia, and whether the patient has a normal ECG. From these 345,779 ECGs exams, only 134,657 are labeled as normal ECGs. Lastly, if the patient died during the follow-up period, the metadata contains the time of the death of the patient. The mean follow-up time is 3.68 years and the mortality rate in this dataset is 3.6%.

2.2. Proposed Method

Our DL approach employs a ResNeXt-based one-dimensional architecture [Xie et al. 2017b] and investigates several ECG lead-set configurations in both tasks of age and sex prediction. Firstly, we provide a description of the lead sets used in our work, including the total number of leads and the number of independent leads. Next, we present our proposed neural network architecture and describe our training strategy. We then discuss our evaluation approach and provide information on our experimental setup. Figure 1 displays the summary of our proposed methodology.

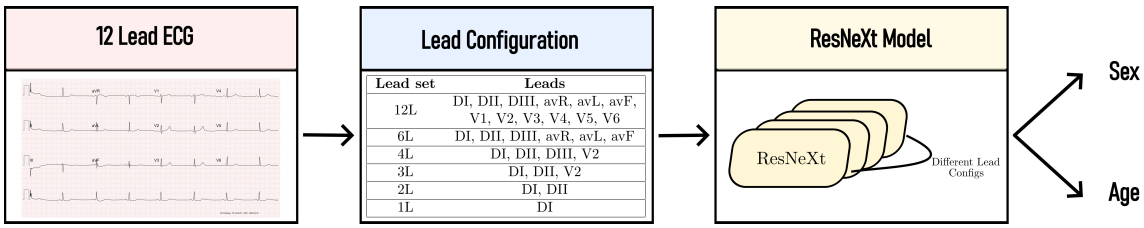

Figure 1. Overview of the proposed method for age and sex estimation from reduced lead ECGs.

2.2.1. Lead configurations

The ECG exam typically employs 12 leads across two distinct planes: frontal (DI, DII, DIII, avR, avL, avF) and transverse (V1, V2, V3, V4, V5, V6). Although the frontal plane consists of 6 leads, only two are independent, meaning that the remaining leads can be derived through vector combination (e.g., DIII = DII - DI). This indicates that these additional leads do not provide further information concerning the frontal plane. On the other hand, the transverse plane presents 6 independent leads. Consequently, the complete 12L ECG contains a total of $2 + 6 = 8$ independent leads.

Our goal in this work is to demonstrate the feasibility of estimating age and sex from reduced lead ECG sets. To this end, we propose five distinct ECG configurations: 12L, 6L, 4L, 3L, 2L, and 1L. With the exception of the 1L configuration, the remaining configurations are derived from the reduced sets proposed in the Computing in Cardiology/Physionet Challenge 2021 [Reyna et al. 2021]. Since commercial smartwatches, such as the AppleWatch, capture only the lead DI, we employed DI as our 1L configuration. We used the 12L configuration as a reference point to analyze and evaluate the impact of utilizing reduced leads to predict sex and age.

Table 1 presents the proposed reduced configurations, identifying the specific leads they represent and the number of independent leads included in each configuration. As can be observed, the 12L lead set exhibits the highest number of independent leads (8). Following that, the 4L and 3L configurations have three independent leads each. The 6L and 2L setups have two independent leads, while the 1L configuration has only one. It is worth noting that technically the 6L, and 2L configurations are equivalent since they all have the same independent leads. Similarly, the 4L and 3L configurations are also equivalent for the same reason.

Table 1. Lead Set Configurations and Independent Leads.

Lead set	Leads	# independent leads
12L	DI, DII, DIII, avR, avL, avF, V1, V2, V3, V4, V5, V6	8
6L	DI, DII, DIII, avR, avL, avF	2
4L	DI, DII, DIII, V2	3
3L	DI, DII, V2	3
2L	DI, DII	2
1L	DI	1

2.2.2. Neural network architecture and Training

In our study, we used a ResNeXt-based architecture to perform sex and age prediction based on ECG signals. The ResNeXt architecture, which builds upon the ResNet architecture [He et al. 2016] and the Inception architecture [Szegedy et al. 2015], was introduced in [Xie et al. 2017a]. It brings notable enhancements over these models, such as: i) Cardinality: the inclusion of multiple parallel paths within a residual block; and ii) Aggregated transform: it employs the split-transform-merge approach proposed in the Inception architecture without the need to fine-tune as many hyperparameters.

In our proposed network, we used an 8-stage architecture with the following components:

- Stage 1: We used a convolutional block that consists in a 1D convolution, batch normalization, and ReLU activation. The block employed a 7x1 filter with 64 filters and a stride of 2.
- Stage 2: A 3x1 max-pooling stage with a stride of 2 was applied.
- Stages 3, 4, 5, and 6: These stages consisted of residual blocks. Each block contained two convolutional blocks with a filter size of 3x1. In the second convolutional block, we employed a cardinality of 8. The number of filters in each residual block doubled at each stage, starting with 64. The components of these are repeated 2 times.
- Stage 7: A global average pooling stage was used followed by a fully connected layer with 1000 units.
- Stage 8: This stage involved an activation function that varied depending on the specific model: ReLU for age estimation and Sigmoid for sex estimation.

Our experiments were carried out using Keras API (version 2.4.3) with TensorFlow backend (version 2.3.0) in Python (version 3.6.8) to build our model. We employed the Adam optimizer with default parameters for training to minimize the cross-entropy loss, a batch size of 128, and a maximum amount of epochs of 30. To avoid overfitting, an early stopping callback with patience of seven epochs was employed.

We partitioned the CODE15 dataset into a training set (70%) and a hold-out test set (30%). In the training set, we performed a 5-fold cross-validation, training an age and sex estimation model for each fold. As a result, we obtained five models for each task. To evaluate these models, we applied them to the hold-out test set and reported the mean and standard deviation of the metrics obtained for each model. It is important to say that in the data splitting process, we took measures to prevent data leakage by ensuring that exams from the same patient did not appear in different splits.

All these steps were carried out using a computer server with four 16 GB V100 GPUS, 346 GB of RAM, and 16 4 GHz CPUs.

2.2.3. Evaluation

We aim to evaluate our method on two tasks: i) sex prediction; and ii) age prediction. To evaluate our method on the first task, we used F1-score (F1), Sensibility (Se), Positive Predictive Value (PPV), Area Under the receiver operating characteristic Curve (AUC). On the second task, we employed the Mean Absolute Error (MAE), Pearson Correlation (ρ), and the Coefficient of Determination (R^2).

We also compared the obtained results between the different lead configurations (12L, 6L, 4L, 3L, 2L, and 1L). To make this comparison, we used a paired t-test between the best performance between all lead configuration and the remaining ones. We used MAE and F1-score to determine the best performing configuration for age and sex prediction, respectively. We considered statistical significance when p-value$<$0.05.

Later, we performed an assessment of the Hazard Ratio (HR) based on the models' misprediction and compared the results obtained for all lead configurations. The HR was calculated using the Cox's proportional hazard model [Cox 1972] adjusted for age, sex, and ECG normal diagnosis. The Cox model is a statistical technique used to examine the relationship between patient survival and various external factors that could impact the outcome. This metric was calculated in three scenarios as follows:

i) The age regression models predict the subject's age to be greater than their actual age by a margin higher than the mean absolute error (MAE) of the entire validation set ($age_{pred} - age_{label} > MAE$);
ii) The age regression models predict the subject's age to be lower than their actual age by a margin higher than the MAE of the entire validation set ($age_{label} - age_{pred} > MAE$).
iii) The sex models make an incorrect prediction of the subject's sex ($sex_{pred} \neq sex_{label}$).

The idea behind the first two scenarios was to verify if there is an increase, decrease or no effect on the mortality when our models' predictions deviate upwards or

downwards from the true subject age by at least a threshold equal to the MAE of the validation set. As for the last one, the idea was to verify if there is an increase, decrease or no effect on mortality when our models predicted a sex different from the label. Notice that hazard ratios greater than 1 indicate an increased risk compared to the reference group, while hazard ratios lower than 1 indicate a decreased risk compared to the reference group.

3. Results

We assessed the performance of various lead sets (12L, 6L, 4L, 3L, 2L, and 1L) in determining age and sex based on ECG data. Tables 2 to 7, summarize the obtained results. The evaluation of our reported results was conducted on a separate 30% hold-out test set. To ensure reliability, the test set was assessed using five models that were trained during cross-validation on the training set. Consequently, we present the mean and standard deviation of the obtained metrics for these five models. Also, we used paired t-tests to assess the statistical differences between different lead sets, using F1 for comparing sex prediction models and MAE for comparing age prediction models.

Table 2 shows the performance metrics for each lead set for predicting sex. The result indicate that the 12L lead set had the highest performance with F1-score of 0.800, Se of 0.807, PPV of 0.793, and AUC of 0.910, with a significant statistical difference between this lead set and the others ($p < 0.05$). After the 12L lead set, the 4L and 3L configurations showed the best performance, with F1-scores of 0.765 and 0.751, respectively. There was no statistically significant difference between the configurations 4L and 3L ($p = 0.28$). In the following, we have lead sets 6L and 2L having F1 of 0.706 and 0.700, respectively. They also do not show statistical differences between each other ($p = 0.58$). The 1L lead set had the lowest performance with F1-score of 0.596.

Similarly, in Table 3, we also show the performance metrics for each lead set in predicting sex but considering only subjects with normal ECGs. By using only ECGs labeled as "normal", we aim to mitigate any possible bias related to associated cardiovascular conditions. The results are similar to those in Table 2, with the 12L lead set having the better results compared to using all patients from the test set (F1 0.833, Se 0.840, PPV 0.793, AUC 0.910), and the 1L lead set having the lowest performance. Overall, the results obtained in these experiments presented better results in all lead set configurations compared to using all patients.

Table 2. Sex-prediction performance metrics for each lead set (all patients)

Leads	F1	Se	PPV	AUC	p-value
12L	**0.800 (0.007)**	**0.807 (0.016)**	**0.793 (0.022)**	**0.910 (0.007)**	**-**
6L	0.706 (0.019)	0.761 (0.094)	0.672 (0.057)	0.842 (0.004)	$<$0.005
4L	0.751 (0.025)	0.780 (0.103)	0.744 (0.075)	0.888 (0.004)	$<$0.005
3L	0.765 (0.011)	0.762 (0.053)	0.774 (0.042)	0.892 (0.004)	$<$0.005
2L	0.700 (0.014)	0.696 (0.066)	0.713 (0.047)	0.839 (0.007)	$<$0.005
1L	0.596 (0.119)	0.580 (0.208)	0.703 (0.089)	0.800 (0.014)	$<$0.005

Additionally, Table 4 presents the performance metrics for each lead set for our age prediction models. The results indicate that the 4L lead set had the lowest MAE of 8.961 years, with ρ of 0.810, and R^2 of 0.637. However, no statistical difference

Table 3. Sex-prediction performance metrics for each lead set (ECG Normal)

Leads	F1	Se	PPV	AUC	p-value
12L	**0.833 (0.005)**	**0.840 (0.014)**	**0.827 (0.022)**	**0.938 (0.004)**	**-**
6L	0.722 (0.016)	0.776 (0.098)	0.692 (0.072)	0.874 (0.004)	<0.005
4L	0.777 (0.027)	0.802 (0.102)	0.775 (0.086)	0.919 (0.003)	<0.005
3L	0.798 (0.009)	0.782 (0.050)	0.820 (0.044)	0.923 (0.003)	<0.005
2L	0.722 (0.008)	0.714 (0.059)	0.739 (0.049)	0.870 (0.006)	<0.005
1L	0.621 (0.101)	0.604 (0.197)	0.719 (0.097)	0.832 (0.015)	<0.005

was found between 12L (MAE 9.021) and the 4L lead set (p = 0.806). The 3L (MAE 9.587) and the 4L lead set showed no statistical difference as well (p = 0.08). After these lead sets, the best performing model is the 2L (MAE 10.569) followed by the 1L (MAE 11.631). Likewise, Table 5 presents the performance metrics for each lead set in predicting age using ECG data from subjects with normal ECGs. Similar to Table 4, the 4L lead set configuration had the lowest MAE and the 1L lead set the highest MAE.

Table 4. Age-prediction performance metrics for each lead set (all patients)

Leads	MAE	ρ	R^2	p-value
12L	9.021 (0.497)	0.811 (0.008)	0.637 (0.036)	0.806
6L	9.795 (0.614)	0.780 (0.010)	0.569 (0.058)	0.019
4L	**8.961 (0.180)**	**0.810 (0.004)**	**0.637 (0.014)**	**-**
3L	9.587 (0.686)	0.797 (0.012)	0.587 (0.052)	0.08
2L	10.569 (1.435)	0.775 (0.018)	0.511 (0.129)	0.038
1L	11.631 (0.328)	0.679 (0.021)	0.414 (0.034)	<0.005

Table 5. Age-prediction performance metrics for each lead set (ECG Normal)

Leads	MAE	ρ	R^2	p-value
12L	8.624 (0.547)	0.814 (0.010)	0.639 (0.044)	0.633
6L	9.608 (0.793)	0.777 (0.008)	0.551 (0.072)	0.015
4L	**8.497 (0.169)**	**0.812 (0.005)**	**0.642 (0.014)**	**-**
3L	9.225 (0.712)	0.797 (0.013)	0.587 (0.057)	0.057
2L	10.247 (1.324)	0.773 (0.017)	0.501 (0.124)	0.019
1L	11.627 (0.413)	0.651 (0.026)	0.362 (0.045)	<0.005

Table 6 shows the hazard ratios for each scenario of the mispredictions of the models trained with different lead sets. For scenarios i) and ii), we used a MAE value corresponding to the MAE obtained in the validation set for each lead set: (12L: MAE 8.9, 6L: MAE 9.8, 4L: MAE 8.9, 3L: MAE 9.5, 2L: MAE 10.5, 1L: MAE 11.6). The hazard ratios were calculated using the Cox's regression model adjusted for age, sex, and normal ECG. It can be seen in this table that there is an increased hazard ratio in scenarios i) and iii) regardless of the lead set. On the other hand, in scenario ii), the hazard ratio decreased for every lead set. Moreover, Table 7 displays the same scenarios as Table 6, but using only patients with normal ECGs. We observed a similar behavior as in Table 7 with increased hazard ratios for all leads in scenarios i) and iii) and a decrease in hazard ratios for scenario ii).

Table 6. Hazard ratios and p-values for different lead sets according to the differences between predicted and true label age along with the differences between predicted and true label sex (all patients).

	Age - Label > MAE		Age - Label < -MAE		Sex ≠ Label	
Leads	*HR(CI 95%)*	*p-value*	*HR(CI 95%)*	*p-value*	*HR(CI 95%)*	*p-value*
12L	2.49 (2.17-2.87)	<0.005	0.71 (0.64-0.77)	<0.005	1.36 (1.24-1.49)	<0.005
6L	2.17 (1.88-2.50)	<0.005	0.71 (0.64-0.78)	<0.005	1.26 (1.15-1.37)	<0.005
4L	2.53 (2.20-2.91)	<0.005	0.74 (0.67-0.81)	<0.005	1.38 (1.26-1.51)	<0.005
3L	2.54 (2.23-2.90)	<0.005	0.71 (0.64-0.78)	<0.005	1.33 (1.22-1.46)	<0.005
2L	2.76 (2.37-3.22)	<0.005	0.75 (0.68-0.82)	<0.005	1.20 (1.10-1.31)	<0.005
1L	2.65 (2.26-3.12)	<0.005	0.64 (0.59-0.71)	<0.005	1.12 (1.03-1.23)	0.01

Table 7. Hazard ratios and p-values for different lead sets according to the differences between predicted and true label age along with the differences between predicted and true label sex (ECG Normal)

	Age - Label > MAE		Age - Label < -MAE		Sex ≠ Label	
Leads	*HR(CI 95%)*	*p-value*	*HR(CI 95%)*	*p-value*	*HR(CI 95%)*	*p-value*
12L	2.24 (1.58-3.17)	<0.005	0.68 (0.54-0.85)	<0.005	1.32 (1.02-1.70)	0.04
6L	1.86 (1.32-2.61)	<0.005	0.62 (0.48-0.80)	<0.005	1.62 (1.31-2.00)	<0.005
4L	1.98 (1.39-2.83)	<0.005	0.74 (0.60-0.93)	0.01	1.40 (1.10-1.77)	<0.005
3L	2.23 (1.60-3.10)	<0.005	0.62 (0.49-0.79)	<0.005	1.64 (1.31-2.07)	<0.005
2L	2.49 (1.74-3.58)	<0.005	0.72 (0.57-0.90)	0.01	1.56 (1.25-1.94)	<0.005
1L	2.01 (1.35-3.01)	<0.005	0.69 (0.55-0.86)	<0.005	1.41 (1.12-1.78)	<0.005

4. Discussion

In this study, we proposed a DL-based approach to predict sex and age from reduced lead ECGs (12L, 6L, 4L, 3L, 2L, and 1L) and evaluated their implications for evaluating patient mortality. The use of reduced lead ECGs presents an opportunity to develop AI techniques that have a broader impact on the general public than the traditional 12L ECG.

In the sex estimation task, our best results were achieved using the 12L approach, which obtained an F1-score of 0.800. We observed comparable results with reduced lead sets, such as 4L and 3L. Following these, the 6L and 2L lead sets obtained F1-scores of 0.706 and 0.700, respectively. The 1L lead set performed the poorest. As expected, the lead sets with more independent leads achieved better results. The 12L configuration has 8 independent leads, while the 4L and 3L configurations, despite having fewer total numbers of leads than 6L, have 3 independent leads. The 1L configuration has the smallest number of independent leads (i.e., 1 independent lead) and also performed the worst.

For age prediction, our best model achieved a mean absolute error (MAE) metric of 8.961 when using the 4L configuration. However, there was no statistically significant difference found between this configuration, the 12L ($p = 0.806$) and 3L ($p = 0.08$) lead sets. These results suggest that we can estimate age using fewer leads (3L and 4L) compared to the standard 12-lead ECG commonly used in age estimation studies [Lima et al. 2021]. The 6L and 2L lead sets, although performing slightly worse than the 4L configuration, achieved comparable results to the 12L configuration. As expected, the 1L lead set exhibited the lowest performance in age prediction.

One could argue that the age and sex prediction models are not directly estimating age or sex, but rather identifying ECG exams with abnormal conditions that may be more prevalent in certain age groups or specific sexes. Therefore, we also evaluated our models using only normal ECG exams. The results of this analysis, presented in Tables 3 and 5, for sex and age estimation, respectively, demonstrate similar behavior between the subset of normal ECGs and the total hold-out test set, which includes all ECGs (shown in Tables 2 and 4).

In summary, the evaluation of different lead-set configurations revealed that even with a reduced number of leads, our models achieved comparable performance to those obtained using the conventional 12-lead ECG setup. This suggests that our approach can be applied in settings where only limited leads are available, such as with devices like the Apple Watch or AliveCor.

Comparing our obtained results with other works in the literature is a difficult task for some reasons. Different studies in the literature have employed diverse datasets, making a direct comparison challenging. [Lima et al. 2021] is the closest related work for comparison but, unfortunately, they only predicted age with ECGs. They proposed an age estimation method from 12L ECGs trained using the CODE dataset. We only used the publicly available version of the CODE dataset which only contains 15% of the data. While they achieved 8.38 for MAE using 12L ECG, we obtained 8.961 and 9.021, respectively, for 4L and 12L. Despite using a significantly smaller training set, our results are comparable to theirs.

Furthermore, our study showed an association between the predictions made by our age estimation and sex identification models and patient mortality for all lead sets, as can be seen in Table 6. When our models predicted ages higher than the actual age by a threshold equivalent to the MAE obtained in the validation set, we observed an increase in mortality. This was measured by the hazard ratio and was consistent across all lead sets. When our models predicted ages lower than the actual age by the same threshold, we found a decrease in mortality. We also demonstrated that misclassifying a subject's sex using our models also led to an increase in mortality rates.

Once again, one can argue that these hazard ratio analyses, although adjusted for normal ECG, can be biased where our models could be making more errors among individuals with diseases and that is the reason they show increased mortality in such cases. Therefore, we performed a hazard ratio analysis only in individuals with normal ECGs, as shown in Table 7. As can be seen in this Table, we observe a similar behavior as Table 6, where there is an increase in mortality among individuals where our models predict the age above the actual age by a certain threshold and when our model misclassifies the sex. Also, there is a decrease in mortality when our age model predicts an age lower than the actual age by a certain threshold. So, even using the 1L lead set, which obtained the lowest results for both sex and age prediction, we were still able to observe significant changes in mortality rates. This could have important clinical applications since 1L devices are increasingly more accessible to the general public through smartwatches such as the AppleWatch. Likewise, 2L and 6L lead sets, which are the ones in devices such as the AliveCor, also provided significant information regarding the mortality rate.

The reasons behind the observed increase in mortality when our age model pre-

dicts values higher than the actual age by a certain threshold, as well as the decrease in mortality when the predicted age is lower than the actual age by the same threshold, require further investigation for a more in depth understanding. One possible interpretation is that during the age model's training phase, it learns the characteristic patterns of ECGs associated with each age group. Therefore, when the model predicts an age for a new ECG and the prediction deviates from the actual age by a certain threshold, it indicates that the ECG exhibits patterns that are different from what is expected for that age group. For ECGs with higher age predictions than the actual age, this could potentially mean an acceleration of aging processes or physiological changes that are more typical of older individuals. On the other hand, ECGs with lower age predictions may suggest a deceleration of aging processes or physiological patterns more commonly observed in younger individuals. These deviations from the expected age-related patterns could be contributing factors to the observed differences in mortality rates among these groups.

A similar interpretation can be applied to the misclassification of sex by our sex model. During training, the model learns the characteristic features of ECGs corresponding to each sex. When the model incorrectly predicts the sex of an ECG, it implies that the ECG exhibits patterns that are outside the expected distribution for that sex. This could potentially indicate underlying hormonal dysfunctions or physiological variations that are not aligned with the typical ECG patterns for the predicted sex. However, these interpretations are speculative, and further detailed investigations are needed to better understand the underlying mechanisms and validate these hypotheses. Future studies should explore the association between these model predictions, age-related physiological changes, sex-related factors, and their potential implications for mortality outcomes.

Our study has some important limitations. Firstly, we only used ECG data from the CODE15 dataset. However, the CODE15 dataset is the largest publicly available ECG dataset to date, which makes it a suitable choice for our study. Also, even though we applied restrictions on the number of leads, it is important to recognize that the CODE15 dataset comprises ECGs captured in resting 12-lead settings. Then, it may not accurately represent ECGs captured in scenarios with limited leads, such as those obtained from devices like the Apple Watch or AliveCor. Nonetheless, we only investigated the impact of predicting age and identifying sex in mortality outcomes. Future studies could explore the potential relationship between model performance for specific diseases and its impact on mortality rates.

5. Conclusion

Our work demonstrates the potential of AI techniques to predict age and identify sex using reduced lead ECGs. The 12L lead set yielded the best results for sex estimation, but 4L and 3L configurations achieved comparable results to the 12L in this task. Additionally, we were able to achieve the best results for sex identification using the 4L lead set. In addition, we were able to gain insights into mortality using our age and sex estimation models for all tested lead configurations. A difference between the predicted and actual age higher than a threshold, as well as a mistake in the sex prediction, led to a higher mortality risk. Also, predicting an age lower than the actual age by a certain threshold yields a lower mortality rate for all lead sets. These findings suggest the feasibility of using age and sex prediction models in limited lead ECG equipment and wearable devices. Nonetheless, with the popularization of devices that provide reduced lead ECGs, such

as the AppleWatch and AliveCor, the development of methods for reduced lead ECGs presents an opportunity to develop AI techniques with a broader impact on the general public, not being restricted to the conventional 12L ECG. Overall, our research highlights the efficacy of reduced lead ECGs in predicting age and sex, and providing valuable insights into mortality.

Acknowledgements

This study was financially supported by Foxconn Brazil and the Zerbini Foundation as part of the research project "Machine Learning in Cardiovascular Medicine".

Competing interests

The authors declare no competing interests.

References

AlGhatrif, M. and Lindsay, J. (2012). A brief review: history to understand fundamentals of electrocardiography. *Journal of community hospital internal medicine perspectives*, 2(1):14383.

Attia, Z. I., Friedman, P. A., Noseworthy, P. A., Lopez-Jimenez, F., Ladewig, D. J., Satam, G., Pellikka, P. A., Munger, T. M., Asirvatham, S. J., Scott, C. G., Carter, R. E., and Kapa, S. (2019a). Age and sex estimation using artificial intelligence from standard 12-lead ecgs. *Circulation: Arrhythmia and Electrophysiology*, 12(9):e007284.

Attia, Z. I., Kapa, S., Yao, X., Lopez-Jimenez, F., Mohan, T. L., Pellikka, P. A., Carter, R. E., Shah, N. D., Friedman, P. A., and Noseworthy, P. A. (2019b). Prospective validation of a deep learning electrocardiogram algorithm for the detection of left ventricular systolic dysfunction. *Journal of cardiovascular electrophysiology*, 30(5):668–674.

Batchvarov, V. N., Ghuran, A., Smetana, P., Hnatkova, K., Harries, M., Dilaveris, P., Camm, A. J., and Malik, M. (2002). Qt-rr relationship in healthy subjects exhibits substantial intersubject variability and high intrasubject stability. *American Journal of Physiology-Heart and Circulatory Physiology*, 282(6):H2356–H2363.

Cohen-Shelly, M., Attia, Z. I., Friedman, P. A., Ito, S., Essayagh, B. A., Ko, W.-Y., Murphree, D. H., Michelena, H. I., Enriquez-Sarano, M., Carter, R. E., et al. (2021). Electrocardiogram screening for aortic valve stenosis using artificial intelligence. *European heart journal*, 42(30):2885–2896.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Lima, E. M., Ribeiro, A. H., Paixão, G. M. M., Ribeiro, M. H., Pinto-Filho, M. M., Gomes, P. R., Oliveira, D. M., Sabino, E. C., Duncan, B. B., Giatti, L., Barreto, S. M., Meira Jr, W., Schön, T. B., and Ribeiro, A. L. P. (2021). Deep neural network-estimated electrocardiographic age as a mortality predictor. *Nat Commun*, 12(1):5117.

Macfarlane, P., McLaughlin, S., Devine, B., and Yang, T. (1994). Effects of age, sex, and race on ecg interval measurements. *Journal of Electrocardiology*, 27:14–19. Research and Technology Transfer in Computerized Electrocardiology.

Macfarlane, P. W. and Kennedy, J. (2021). Automated ecg interpretation—a brief history from high expectations to deepest networks. *Hearts*, 2(4):433–448.

Malik, M., Hnatkova, K., Kowalski, D., Keirns, J. J., and van Gelderen, E. M. (2013). Qt/rr curvatures in healthy subjects: sex differences and covariates. *American Journal of Physiology-Heart and Circulatory Physiology*, 305(12):H1798–H1806.

PIPBERGER, H. V., FREIS, E. D., TABACK, L., and MASON, H. L. (1960). Preparation of electrocardiographic data for analysis by digital electronic computer. *Circulation*, 21(3):413–418.

Rafie, N., Kashou, A. H., and Noseworthy, P. A. (2021). Ecg interpretation: Clinical relevance, challenges, and advances. *Hearts*, 2(4):505–513.

Reyna, M. A., Sadr, N., Alday, E. A. P., Gu, A., Shah, A. J., Robichaux, C., Rad, A. B., Elola, A., Seyedi, S., Ansari, S., Ghanbari, H., Li, Q., Sharma, A., and Clifford, G. D. (2021). 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.

Ribeiro, A. H., Ribeiro, M. H., Paixão, G. M., Oliveira, D. M., Gomes, P. R., Canazart, J. A., Ferreira, M. P., Andersson, C. R., Macfarlane, P. W., Meira Jr, W., et al. (2020). Automatic diagnosis of the 12-lead ecg using a deep neural network. *Nature communications*, 11(1):1760.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9.

Xie, S., Girshick, R., Dollár, P., Tu, Z., and He, K. (2017a). Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1492–1500.

Xie, S., Girshick, R., Dollár, P., Tu, Z., and He, K. (2017b). Aggregated residual transformations for deep neural networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5987–5995.