Tabea F. A. Steinbrinker*, Philip Hempel, Dagmar Krefting, and Nicolai Spicher

Explaining a Sex-Related Cardiovascular Risk Continuum

https://doi.org/10.1515/cdbme-2025-0179

Abstract: Cardiovascular diseases (CVDs) remain the leading cause of mortality worldwide. New sensor-based approaches to assess cardiovascular risk may support individualized healthcare. Recent advancements in Artificial Intelligence (AI) suggest Electrocardiogram (ECG) - based predictions of age and sex as potential biomarkers for cardiovascular risk stratification. Discrepancies between predicted and chronological age, as well as inconsistencies in sex classification, have been linked to increased cardiovascular risk. However, these data-driven approaches also contain the risk of not equally performing across subgroups. In this work, we introduce a deep learning model that predicts both ECG-derived age and sex simultaneously. Trained on the CODE dataset and validated on PTB-XL, the model achieves a mean absolute error of 8.85 years in age prediction and an Area under ROC of 0.93 for sex classification. Notably, sex prediction accuracy decreases with age, suggesting morphological changes in ECG signals. To enhance interpretability, we use posthoc explainable AI analysis to highlight ECG regions relevant for sex classification, revealing known sex-specific features. Our findings underline the potential of AI-driven ECG analysis as a promising, explainable tool for cardiovascular risk assessment.

Keywords: ECG, AI, Sex-specific medicine, Cardiovascular risk

1 Introduction

Cardiovascular diseases (CVDs) remain the leading cause of mortality worldwide, necessitating early and accurate risk assessment strategies. Traditional cardiovascular risk scores incorporate well-established factors such as blood pressure, cholesterol levels, smoking status, and family history. However, emerging evidence suggests that applying Artificial Intelligence (AI) models directly to a raw electrocardiogram (ECG) may serve as novel physiology-based biomarkers for cardiovascular risk stratification.

Recent advances in deep learning have enabled the estimation of ECG-derived age and sex prediction, both of which have been shown to reflect underlying cardiovascular health. Discrepancies between predicted and chronological age larger than the models mean average error (MAE) of about 8 years could be related to different outcomes, such as mortality or the development of congestive heart failure. Underestimation and overestimation of the age were associated with better and worse clinical outcomes, respectively [1, 2]. The definition of a sex discordance score, a continuous variable ranging from 0 (predicted male sex) to 1 (predicted female sex) allowed for the investigation of sex related cardiovascular risk. A lower discordance could be shown in persons of age over 40 years, and in particular ECGs showing tachycardia (high heart rate) or broad QRS complexes. Interestingly, sex discordance in women with a normal ECG in an outpatient setting was associated with higher cardiovascular risk, but not in male [3]. These findings suggest that discordance analysis of ECG-based sex and age predictions may serve as an individualized risk assessment related to deviations from expected cardiac physiology.

At the same time, AI-based approaches in general potentially suffer from a lack of trustworthiness including explainability [4]. Due to the high dimensionality of ECG data, training biases such as device-specific artefacts in a certain group of interest may be present, even if the training set appears to be balanced regarding sensitive variables, such as age, sex, race/ethnicity, or disability status [5].

In this work, we propose an extension of a state-of-the-art AI model developed for age prediction, which now enables to simultaneously predict age and sex from standard diagnostic ECG. Furthermore, we analyze the models explainability, by using different methods including explainable AI analysis.

2 Material and Methods

2.1 Data

Two different datasets are used for training and validation.

***Corresponding author: Tabea F. A. Steinbrinker,** Department of Medical Informatics, University Medical Center Göttingen, Robert-Koch-Straße 40, 37075 Göttingen, Germany, e-mail: tabea.steinbrinker@med.uni-goettingen.de
Philip Hempel, Dagmar Krefting, Department of Medical Informatics, University Medical Center Göttingen, Robert-Koch-Straße 40, 37075 Göttingen, Germany
Nicolai Spicher, Department of Medical Informatics, University Medical Center Göttingen, 37075 Göttingen, Germany and Department of Health Technology, Technical University of Denmark, 2800 Kgs. Lyngby, Denmark

Training dataset – CODE The Brazilian Clinical Outcomes in Digital Electrocardiography (CODE) dataset was utilized [6, 7], containing clinical 12-lead ECGs of 10 second length with annotations for various abnormalities. The dataset was collected from $1,676,384$ patients older than 16 years in the Telehealth Network of Minas Gerais, Brazil between 2010 and 2016. The patients in the dataset had a mean age of 51.2 years and 60.3% were female.

Validation dataset – PTB-XL For testing and further analysis of the model the PTB-XL dataset was used [8, 9]. It consists of $21,799$ clinical 12-lead ECGs of 10 second length from $18,869$ patients, collected between 1989 and 1996 with devices of the Schiller AG. We exclude all patients that are < 20 and > 85 years of age, resulting in $20,370$ ECGs (mean age: 59.64 years, median age: 61 years, 46.64% female).

2.2 AI Model, Validation, and XAI analysis

We modified the 1DResNet architecture developed for age prediction [1] to simultaneously perform a sex and an age prediction. To achieve this, before the last fully connected layer of the model, it was split into two "heads". The age prediction head employed a fully connected linear layer and outputs a scalar value, as it was implemented in the original model. For the sex prediction head, a further multi layer perceptron was included. It contains a hidden layer of 128 nodes, followed by a ReLu activation function. Finally, a single scalar value was outputted and interpreted as a probability distribution with a sigmoid layer. Figure 1 gives a visualization of these two heads. To consider the different scales of age and sex, the loss function was altered to include a relative error of the age, as well as the binary cross entropy (BCE) for the sex, i.e.

$$\text{Loss} = \text{loss}_{\text{age}} + \text{loss}_{\text{sex}},$$

where

$$\begin{aligned}
\text{loss}_{\text{age}} &= \left(\frac{\text{age}_{\text{true}} - \text{age}_{\text{pred}}}{\text{age}_{\text{pred}}} \right)^2, \\
\text{loss}_{\text{sex}} &= \text{BCE}(\text{sex}_{\text{true}}, \text{sex}_{\text{pred}}) \\
&= -\text{mean}\left(\left[\text{sex}^n_{\text{true}} \cdot \log(\text{sex}^n_{\text{pred}}) \right.\right. \\
&\qquad \left.\left. + \ (1 - \text{sex}^n_{\text{true}}) \cdot \log(1 - \text{sex}^n_{\text{pred}}) \right]_{n=0}^{N-1} \right)
\end{aligned}$$

and $\text{sex}^n_{\text{true}}$ and $\text{sex}^n_{\text{pred}}$ are the true value and prediction of the n-th sample, respectively. The used implantation of PyTorch BCELoss function clamps its log function outputs to be greater than or equal to -100, to ensure the output to be finite, even though $\log(0) = -\infty$.

The sex prediction is given as a value between 0 and 1, with 0 indicating the prediction of a male and 1 the prediction of a female. The sex prediction will be treated as a continuous variable.

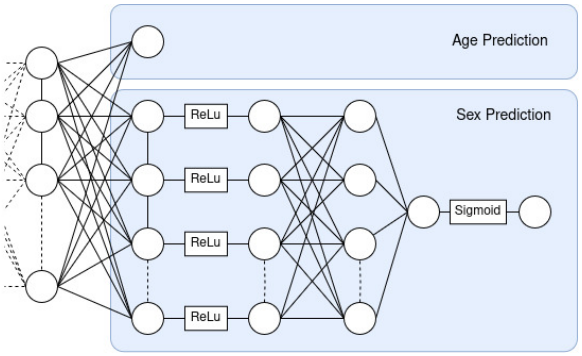

Fig. 1: Visualization of the two "heads" of the model. It connects to the last fully connected layer of the ResNet Model. The upper part outputs the age prediction, the lower part outputs the sex prediction.

The patients are separated into three age groups: below 40, between $40 - 60$ years, and above 60 years. These boundaries are chosen as they represent changes in the human hormone levels, as happening in menopause [10].

The XAI method integrated gradients (IG) [11] validated in its use for ECGs [12] was used to determine which parts of the ECG where most relevant for the sex prediction. For this, "relevances" are computed for each input sample of each processed ECG. To visualize these relevances, an existing pipeline [2] was used: The freely available QRS detector xQRS [13] was used to identify R-peaks and fixed-length segments of the relevances of 650 ms around the R-Peak where stored. These relevances where normalized and averaged across all $20,370$ ECGs.

3 Results

Analyzing the ROC curve in Figure 2 demonstrates that a threshold of 0.357 maximizes the F1-Score for the sex prediction and achieves a value of 0.84. The Area under the ROC curve is 0.93.

Figure 3 shows the age predictions on the PTB-XL dataset. The model shows a mean absolute error (MAE) of 8.85 years and Pearsons r-value of 0.79. Age predictions range from 20 to 100 years, with visible underestimations and overestimations in all chronological ages. The distributions of both the chronological and the predicted age show a maximum around 65 years. Figure 4 shows the histogram of the sex prediction with red color indicating a male biological sex and turquoise indicating female biological sex. A clear difference between both histograms can be observed.

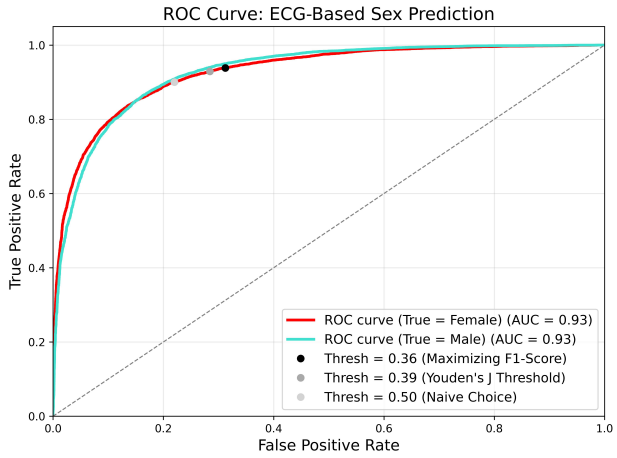

Fig. 2: ROC curve: Classification on ECGs from females (red curve) and males (turquoise curve) with three different thresholds (gray dots)

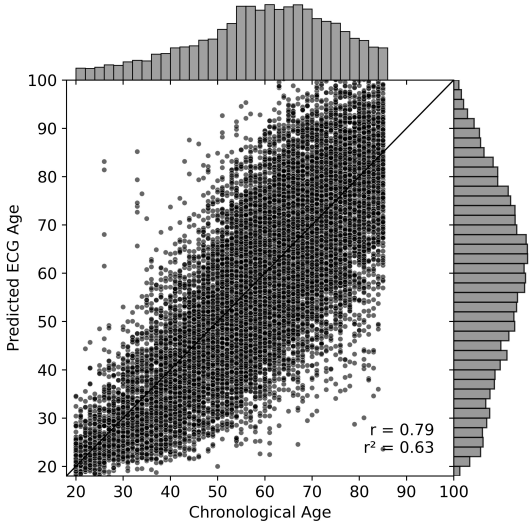

Fig. 3: Age prediction: The ECG-Age predictions for all subjects in the PTB-XL dataset given as a function of the subject's chronological age. The histograms show the distribution of the chronological age on the top and of the predicted ECG-Age at the right side, respectively.

We observe, that the uncertainty of the sex prediction, i.e. a prediction diverging from 0 for male and 1 for female patients, increases with chronological age.

Figure 5 shows the distribution of the continuous sex prediction value for the three different age groups. A highly significant increase in the mean error between these age groups can be observed: all distributions obtain a p-value by a two sided t-test of $p < 0.001$.

In Figure 6 the averaged relevances of lead V1 are shown. As can be seen, there is an area of high positive relevances at the end of the QRS complex (≈ 350ms).

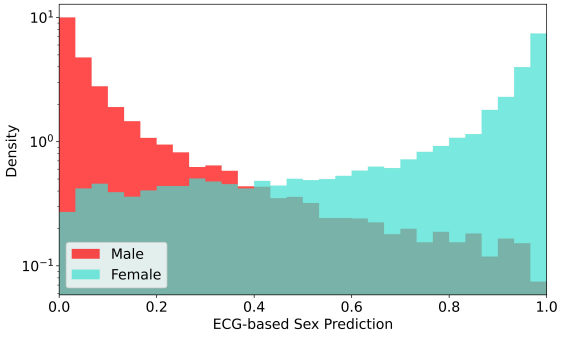

Fig. 4: Sex prediction: Logarithmic histogram of the sex predictions for all subjects of PTB-XL. Male sex is assigned to the value 0, and female sex to the value 1, the predictions reflect the probability of female sex.

4 Discussion

Although the model can not only predict the age but also the sex from an ECG, we observe a similar performance as previously trained models predicting age only [6]. Furthermore, we achieve similar performance to previous models only predicted sex (our AUC 0.93 vs a AUC of 0.94 in [3]). Results clearly show, that the sex prediction is significantly better, when patients are younger. This supports the strong role that sex hormones have on the ECG morphology, that become visible during puberty and progressively subside with advancing age [14]. For females, the main sex hormone is estrogen, where a strong decrease is experienced during menopause, happening at a mean age of around 50 years [10]. This may be an explanation in the great difference of prediction quality of the under 40 year and over 60 age groups.

The XAI analysis further support these findings. While in previous work, age prediction only has been analyzed [15], our new findings suggests that the model learned known features

The

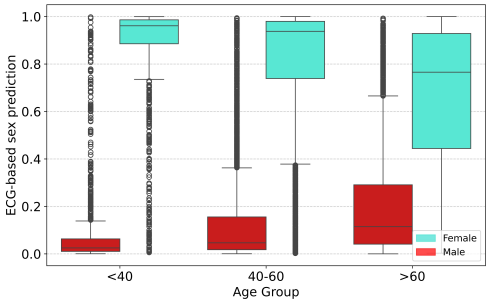

Fig. 5: Sex prediction: Prediction error of the sex prediction divided by three age groups. The boxes show the 50% percentile, the lower whiskers shows the $Q1 - 1.5 * IQR$ and the upper whisker shows $Q3 + 1.5 * IQR$, where IQR is the inter quartile range. The line in the box indicates the median.

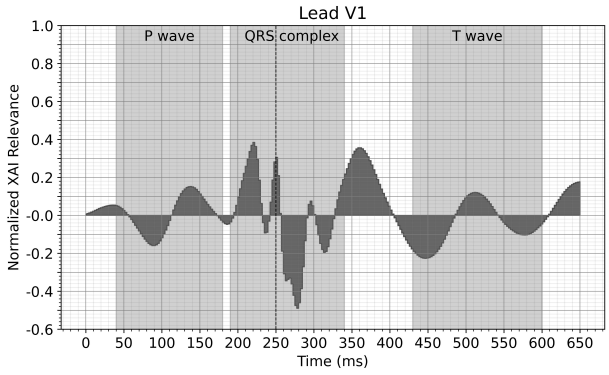

Fig. 6: XAI for sex prediction: IG relevances normalized over all subjects for lead V1.

in differences between male and female ECGs. As shown in Figure 6, the model has a focus on the ST-elevation, i.e. high relevances right after the end of the QRS complex and before the beginning of the T-wave, an ECG feature which is reported in around 90% of male ECG but only 20% of female ECG [14]. In general, woman have narrower QRS complexes than men. The relevances indicate this in the great negative relevance at the S-wave of the ECG.

5 Conclusion

ECG-based AI models are currently intensively researched as potential risk assessment tools. By including sex into the model predictions, we aim to enhance their explainability towards this sensitive variable, while simultaneously pointing out potential differences in male and female ECGs.

Code Availability

The code for training and evaluation can be found here.

Author Statement

This project was partly funded by the German Federal Ministry of Education and Research (ACRIBiS grant n. 01ZZ2317B) as well as by the Lower Saxony "Vorab" of the Volkswagen Foundation and the Ministry for Science and Culture of Lower Saxony, Germany under Grant 76211-12-1/21. Authors state no conflict of interest.

References

[1] E. M. Lima, A. H. Ribeiro, G. M. Paixão, M. H. Ribeiro, M. M. Pinto-Filho *et al.*, "Deep neural network-estimated electrocardiographic age as a mortality predictor," *Nature Communications*, vol. 12, no. 1, p. 5117, 2021.

[2] P. Hempel, A. H. Ribeiro, M. Vollmer, D. Krefting, N. Spicher *et al.*, "Explainable ai associates ecg aging effects with increased cardiovascular risk in a longitudinal population study," *npj Digital Medicine*, vol. 8, no. 1, p. 25, 2025.

[3] A. Sau, E. Sieliwonczyk, K. Patlatzoglou, L. Pastika, K. A. McGurk *et al.*, "Artificial intelligence-enhanced electrocardiography for the identification of a sex-related cardiovascular risk continuum: A retrospective cohort study," *The Lancet Digital Health*, vol. 7, no. 3, Mar 2025.

[4] B. Li, P. Qi, B. Liu, S. Di, J. Liu *et al.*, "Trustworthy AI: From principles to practices," vol. 55, no. 9, pp. 177:1–177:46. [Online]. Available: https://dl.acm.org/doi/10.1145/3555803

[5] M. Liu, Y. Ning, S. Teixayavong, M. Mertens, J. Xu *et al.*, "A translational perspective towards clinical AI fairness," *npj Digital Medicine*, vol. 6, no. 1, p. 172, Sep. 2023. [Online]. Available: https://www.nature.com/articles/s41746-023-00918-4

[6] A. L. P. Ribeiro, G. M. Paixao, P. R. Gomes, M. H. Ribeiro, A. H. Ribeiro *et al.*, "Tele-electrocardiography and bigdata: the code (clinical outcomes in digital electrocardiography) study," *Journal of electrocardiology*, vol. 57, pp. S75–S78, 2019.

[7] A. H. Ribeiro, M. H. Ribeiro, G. M. Paixão, D. M. Oliveira, P. R. Gomes *et al.*, "CODE dataset," *Dataset*, 11 2021. [Online]. Available: https://figshare.scilifelab.se/articles/dataset/CODE_dataset/15169716

[8] P. Wagner, N. Strodthoff, R.-D. Bousseljot, W. Samek, and T. Schaeffter, "Ptb-xl, a large publicly available electrocardiography dataset (version 1.0.3)," 2022. [Online]. Available: https://doi.org/10.13026/kfzx-aw45

[9] P. Wagner, N. Strodthoff, R.-D. Bousseljot, D. Kreiseler, F. I. Lunze *et al.*, "Ptb-xl: A large publicly available ecg dataset," *Scientific Data*, 2020. [Online]. Available: https://doi.org/10.1038/s41597-020-0495-6

[10] E. Von der Lippe and F. Prütz, "Age at natural menopause: Results from the german health interview and examination survey," *European Journal of Public Health*, vol. 26, p. ckw172.078, 11 2016. [Online]. Available: https://doi.org/10.1093/eurpub/ckw172.078

[11] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," *CoRR*, vol. abs/1703.01365, 2017. [Online]. Available: http://arxiv.org/abs/1703.01365

[12] T. Bender, J. M. Beinecke, D. Krefting, C. Müller, H. Dathe *et al.*, "Analysis of a deep learning model for 12-lead ecg classification reveals learned features similar to diagnostic criteria," *IEEE Journal of Biomedical and Health Informatics*, vol. 28, no. 4, p. 1848–1859, Apr 2024.

[13] G. Moody, T. Pollard, and B. Moody, "Wfdb software package (version 10.7.0)," PhysioNet, 2022, https://doi.org/10.13026/gjvw-1m31.

[14] V. Carbone, F. Guarnaccia, G. Carbone, G. B. Zito, U. Oliviero *et al.*, "Gender differences in the 12-lead electrocardiogram: clinical implications and prospects," *The Italian Journal of Gender-Specific Medicine*, vol. 6, no. 3, pp. 126–141, 9 2020. [Online]. Available: http://dx.doi.org/10.1723/3432.34217

[15] P. Hempel, T. Bender, K. Gandhi, and N. Spicher, "Towards explaining deep neural network-based heart age estimation," in *2023 IEEE EMBS Special Topic Conference on Data Science and Engineering in Healthcare, Medicine and Biology*, 2023, pp. 41–42.