Prediction of Sex From a 12-lead ECG Using Deep Learning: External Validation and Interpretation

Myrte Barthels[1,2,3], Henri Gruwez[1,4,5], Thomas De Cooman[3], Noella Pierlet[1], David Duncker[6], Pieter Vandervoort[1,2,4]

[1] Doctoral School of Medicine and Life Sciences, Hasselt University, Hasselt, Belgium
[2] Department Future Health, Hospital East-Limburg, Genk, Belgium
[3] Qompium NV, Hasselt, Belgium
[4] Department of Cardiology, Hospital East-Limburg, Genk, Belgium
[5] Department of Cardiovascular sciences, University of Leuven, Leuven, Belgium
[6] Department of Cardiology and Angiology, Hannover Medical School, Hannover, Germany

Abstract

Sex-related information might be encoded in the ECG given the inherent biological distinctions between sexes. The aim of this study was to predict sex from the ECG and to uncover the representation of sex within the morphology of the ECG waveform using saliency maps. In a secondary analysis, different ECG rhythms and age groups were compared to examine the effect of variation in the ECG morphology on the outcome. 150052 patients were included with a mean age of 59.2 ± 16.5 years of which 48% were females. The overall model resulted in an accuracy of 0.89. Saliency maps highlighted different parts of the P-wave and the ST-segment across leads as the most predictive regions. Additionally, subgroup analysis revealed decreased performance in subgroups with an altered P-wave and/or ST-segment. This study demonstrates that sex can be predicted from a 12 lead ECG with high accuracy and indicates that sex information is concealed in the P- wave and the ST-segment.

1. Introduction

Electrocardiography (ECG) is used as a non-invasive assessment of the cardiac condition of an individual. Advancements in AI have shown a greater potential of ECG signals beyond its current clinical use and highlight its relevance as a marker for health assessment and characterization of patients [1].

Recent studies have demonstrated that patient characteristics such as age and sex can be predicted from the ECG. Particularly, deep learning (DL) models have achieved remarkable accuracy rates of up to 90% for sex prediction [1–3], a task that is very difficult for cardiologists. The literature attributes this high accuracy to factors such as differences in hormonal status, aging processes, anatomy and biology that are subtly reflected in the ECG signal [1, 4, 5].

Despite these promising outcomes, DL algorithms are often criticized for their "black box" nature, making it difficult to understand the underlying decision-making processes. Recent advancements in explainable AI have introduced techniques such as saliency maps, which help to clarify these processes by highlighting the importance of various input features to the model's predictions [6]. Likewise, saliency maps can be used for ECG models to identify the most predictive regions in an ECG waveform. Examining variations in ECG morphology within these regions could provide deeper insights into the robustness and limitations of these models.

In this paper, we propose a DL sex prediction model and provide insights into the encoding of sex at the signal level using saliency maps. Resulting maps are further explored with a subgroup analysis to examine the effect of variations in the ECG morphology on the model's performance.

2. Materials and methods

2.1. Data

This study included all ECGs from patients older than 18 years with at least one 10-second 12 lead ECG obtained at Ziekenhuis Oost- Limburg (ZOL) (Genk, Belgium) between October 1, 2002 and August 31, 2021. The raw ECG data, along with ECG parameters and quality and diagnostic labels generated by the GE Marquette software, were extracted from the MUSE data management system. All ECGs were acquired at a sampling rate of 500 Hz. ECGs were excluded if there was missing data regarding the sex, age or acquisitiondate. Furthermore, we excluded

ECGs with a bad quality label (electrode artifacts, muscle tremor or baseline wander, muscle tremor and low quality). Patients were randomly split into training (48%), validation (16%) and test (36%) sets. For patients with multiple ECGs over time, only the first available ECG was selected. This study was reviewed and approved by the medical ethical committee of ZOL.

2.2. Model

The deep learning architecture was previously described by Attia et al. [3] and consists of eight convolutional blocks (1D convolution, batchnormalization, relu activation and max pooling) extracting the temporal features from the long axis of the raw ECG, followed by one convolutional block along the short axis to fuse the data from all the ECG leads. Next, two fully connected blocks (fully connectected, batchnormalization, relu activation and dropout) are added and lastly the output is obtained using softmax activation. The final network is trained to predict the sex of the patient and reports a probability of being female as a continuous number in the range 0 to 1. The model input consists of the raw ECG represented in 12 leads (leads I,II, III, AvR, AvL, AvF and V1-V6) and is given by a 12 by 5000 matrix (ECG recorded for 10 s at 500 Hz). The network was implemented using TensorFlow (TensorFlow, Google, Mountain View, CA, USA) backend on Microsoft Azure (Microsoft, Redmond, WA, USA). Hyperparameters were tuned to the data and were finally set at a learning rate of 3e–4, Adam optimizer and batch size of 16.

2.3. Evaluation metrics

The performance of the neural network to differentiate between male and female was assessed using the area under the receiver operating characteristic curve (AUROC), accuracy, specificity, and sensitivity. Two-sided 95% confidence intervals (CI) for the AUROC were estimated using the Sun and Xu optimization of the Delong method [7]. The cut-off threshold for the classification was optimized to the point that yielded similar accuracy, specificity and sensitivity on the validation set.

Saliency maps were calculated over each ECG after which the Pan–Tompkins peak detection algorithm was used to detect the heartbeats [8]. All heartbeats were segmented and aligned on the R peak and an average heartbeat and saliency map was calculated within the ECG and across patients for final visualisation.

In a secondary analysis, we looked into subgroups (age and ECG rhythm) targeted at varying ECG morphology, potentially linking saliency map outcomes to these variations. Both the effect of different ECG rhythms and increasing age were compared using the AUROC metric. Furthermore, we modeled the effect of a correct prediction of the sex on the log probability using a binary logistic regression model. The probability function is given by:

$$P(Correct) = \frac{1}{1 + e^{(\beta_0 + \beta_1 * Age + \beta_2 * AF + \beta_3 * ODR)}}$$

where β_0 is the intercept, β_1 the coefficient of the age represented as the effect of a one year increase in age. β_2 and β_3 are the coefficients representing the change in effect for atrial fibrillation (AF) /atrial flutter (AFl) ECGs and other diverging rhythms (ODR) ECGs, respectively, relative to sinus rhythm (SR) ECGs.

3. Results

3.1. Data

Data extraction from the ECG database identified 218460 patients. After applying the exclusion criteria, we included 150052 patients with mean age of 59.2 $\pm$ 16.5 years of which 48.0% were female. 91.6% of ECGs were acquired in SR, 4.3% were in AF or AFl and 4.2% showed ODR. The algorithm was trained using 71905 ECGs, internally validated using 24022 ECGs and tested on 54125 ECGs. Detailed demographics for the training, validation and testset can be found in Table 1.

Table 1. Dataset demographics. SR = sinus rhythm, AF = atrial fibrillation, AFl = atrial flutter, ODR = other diverging rhythms.

	Patients (%)		
Demographics	**Train**	**Validation**	**Test**
Total number of patients	71905	24022	54125
Sex (Female)	34426 (47.9)	11407 (47.5)	26247 (48.5)
Mean age (years)	59.2 $\pm$16.5	59.2 $\pm$16.5	59.3 $\pm$16.5
Number of AF/AFl ECGs	3070 (4.3)	1035 (4.3)	2279 (4.2)
Number of SR ECGs	65852 (91.6)	21938 (91.3)	49596 (91.6)
Number of ODR ECGs	2983 (4.1)	1049 (4.4)	2250 (4.2)

3.2. Performance

Table 2 shows the performance of the sex classification model. The model obtained an AUROC of 0.956 (95% CI, [0.957-0.960]). Our sex prediction model shows strong performance that is comparable to literature.

Table 2. Model performance as compared to literature. AUROC = area under the receiver operator curve, acc = accuracy, sens = sensitivity, spec = specificity.

Metric	**Our model**	**Attia** [3]	**Demolder** [2]
AUROC	0.956	0.969	0.946
Acc	0.890	0.904	0.880
Sens	0.889	/	0.880
Spec	0.890	/	0.882

Figure 1 shows the saliency map for the average ECG heartbeat in each lead calculated over a random sample of 4096 ECGs in the testset. The results suggest that the area around the P-wave and the ST-segment play an important role in the prediction of the sex.

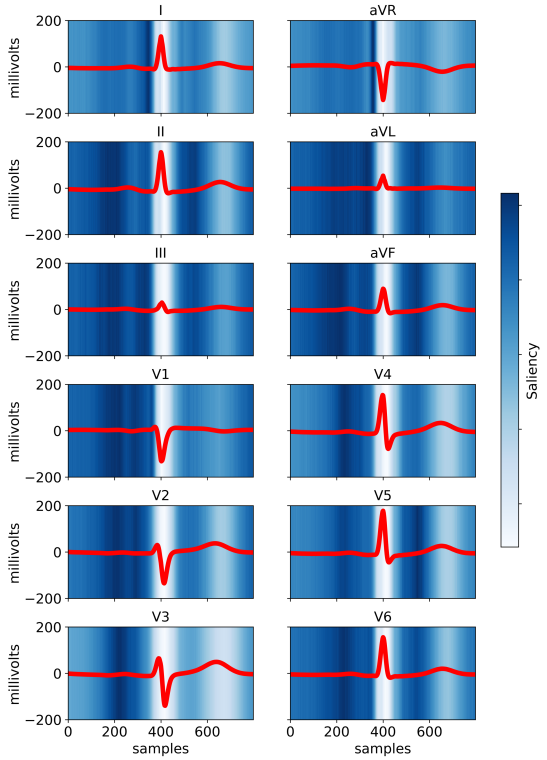

Figure 1. Saliency map of the average ECG heartbeat in each lead. The red line depicts the average heartbeat. The background indicates the gradient. Darker regions indicate higher weight in the prediction.

The results of the secondary analysis are shown in Figure 2. To visualize the effect of the non-categorical value age on the AUROC and the ROC curve, a cutoff at 50 years was chosen. Increased values are observed for patients under 50 years and ECGs in SR (0.97 and 0.98), and decreased values are found for AF/AFl and ODR ECGs as wel as patients over 50 years (0.86, 0.81 and 0.95).

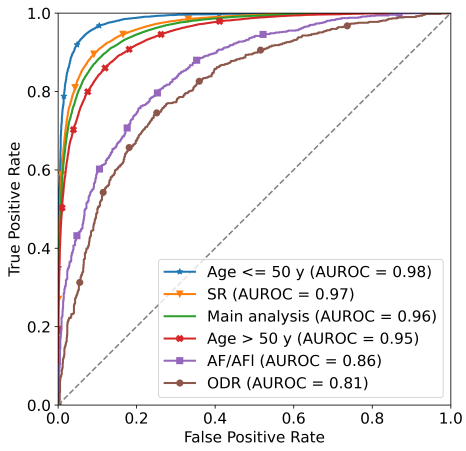

Figure 2. Receiver operator characteristic curve of the main analysis and subgroup analysis. AUROC = area under the receiver operator curve, SR = sinus rhythm, AF = atrial fibrillation, AFl = atrial flutter, ODR = other diverging rhythms, y = years.

Table 3 shows the results of the uni- and multivariate logistic regression. The univariate analysis indicates that age is negatively correlated with the likelihood of a correct prediction. For each additional year of age, the odds of making a correct prediction decrease by about $1-e^{0.030} \approx 3\%$. Even more significant is the impact of the variation in rhythm on the log odds, where, compared to SR, having ODR or AF/AFl decreases the likelihood of a correct prediction by $1 - e^{1.037} \approx 64\%$ and $1 - e^{1.128} \approx 68\%$ respectively. After adjusting for the effect of both variables using a multivariate analysis, age and rhythm remain significant (p-value < 0.001) in decreasing the likelihood of the outcome, with 2.5%, 47% and 56% respectively.

Table 3. Univariate (1) and multivariate (2) logistic regression results. Coeff. = coefficient, CI = confidence interval, SR = sinus rhythm, AF = atrial fibrillation, AFl = atrial flutter, ODR = other diverging rhythms.

	Variable	**Coeff.**	**95% CI**	**p-value**
1	Intercept	3.958	$3.836, 4.079$	< 0.001
	Age	-0.030	$-0.032, -0.028$	< 0.001
1	Intercept	2.216	$2.187, 2.246$	< 0.001
	Rhythm			
	SR-AF/AFl	-1.037	$-1.138, -0.936$	< 0.001
	SR-ODR	-1.128	$-1.228, -1.029$	< 0.001
2	Intercept	3.755	$3.633, 3.878$	< 0.001
	Age	-0.025	$-0.027, -0.024$	< 0.001
	Rhythm			
	SR-AF/AFl	-0.637	$-0.742, -0.532$	< 0.001
	SR-ODR	-0.825	$-0.927, -0.722$	< 0.001

4. Discussion and Conclusion

This study aimed to develop an ECG-based sex prediction model and to explore how sex-related information is encoded in ECG signals. Results indicate a high prediction accuracy of 89.0%. Saliency maps identified the P-wave and the ST-segment as most predictive regions across all leads, suggesting that anatomical and biological differences between the sexes are more prominently pronounced in these regions [1, 4, 5]. However, further investigation is needed to fully understand these findings. Subgroup analysis revealed reduced prediction power for non-SR ECGs and in older individuals. Non-SR ECGs are characterized by irregular electrical activity, which is often reflected in alterations in the P-wave and the ST-segment and may explain the reduced performance in these cases. Furthermore, the age-related decline in performance is partially explained by the higher prevalence of AF and ODR in older populations. However, the multivariate analysis demonstrated that the effect of age remained significant even after correcting for rhythm, as was also found in the study by Lyle et al. [9]. Notably, it was found that ECGs from healthy individuals over 50 years old often show P-wave alterations [10].

Some limitations should be noted. First, saliency maps should be interpreted with care as it is difficult to know whether an explanation is correct. There is a need for evaluation metrics for saliency maps yet current saliency metrics can be statistically unreliable and inconsistent [11–13]. Second, our analysis is performed in a single center, necessitating further validation studies to assess the model's generalizability beyond a Western-European context. Lastly, we want to emphasize that predicting sex from ECG is not clinically crucial, yet, it demonstrates that an ECG contains latent information about a individual's characteristics, challenging to detect with the human eye [1].

In conclusion, sex can be predicted with high accuracy from a 12 lead ECG using deep learning. An AI explainable technique, saliency maps, highlights regions around the P-wave and the ST-segment as most important for its prediction. Subgroup analysis focusing on age and ECG rhythm confirm the significance of these regions in the outcomes. Futher exploration of covariates could help us understand the impact of sex of the ECG.

Acknowledgments

This study is part of Limburg Clinical Research Center, supported by Hasselt University, Ziekenhuis Oost-Limburg and Jessa Hospital.

References

[1] Strodthoff N, Wagner P, Schaeffter T, Samek W. Deep learning for ecg analysis: Benchmarks and insights from ptb-xl. IEEE J Biomed Health Inform May. 2021; 25(5):1519–1528.

[2] Demolder A, Nauwynck M, Pauw MD, Buyzere MD, Duytschaever M, Timmermans F, Pooter JD. Prediction of certainty in artificial intelligence-enabled electrocardiography. J Electrocar Mar. 2024;83:71–79.

[3] Attia ZI, Noseworthy PAFPA, Lopez-Jimenez F, Ladewig DJ, Satam G, Pellikka PA, Munger TM, Asirvatham SJ, Scott CG, Carter RE, Kapa S. Age and sex estimation using artificial intelligence from standard 12-lead ecgs. Circ Arrhythm Electrophysiol Sep. 2019;12(9):e007284.

[4] Naser JA, Lopez-Jimenez F, Chang AY, Baez-Suarez A, Attia ZI, Pislaru SV, Pellikka PA, Lin G, Kapa S, Friedman PA, Noseworthy PA. Artificial intelligence-augmented electrocardiogram in determining sex: Correlation with sex hormone levels. Mayo Clin Proc Apr. 2023;98(4):541–548.

[5] Pierre SRS, Peirlinck M, Kuhl E. Sex matters: A comprehensive comparison of female and male hearts. Circ Arrhythm Electrophysiol Mar. 2022;22(13):831179.

[6] Shrikumar A, Greenside P, Kundaje A. Learning important features through propagating activation differences. ICML Aug. 2017;70:3145–3153.

[7] Sun X, Xu W. Fast implementation of delong's algorithm for comparing the areas under correlated receiver operating characteristic curves. IEE Sign Proc Let Nov. 2014; 21(11):1389–1393.

[8] Pan J, Tompkins WJ. A real-time qrs detection algorithm. IEE Trans Biom Eng Mar. 1985;BME-32(3):220–236.

[9] Lyle JV, Nandi M, Aston PJ. Symmetric projection attractor reconstruction: Sex differences in the ecg. Front Cardiovasc Med Sep. 2021;23(8):709457.

[10] Havmoller R, Carlson J, Holmqvist F, Herreros A, Meurling CJ, Olsson B, Platonov P. Age-related changes in p wave morphology in healthy subjects. BMC Cardiovasc Disord Jul. 2007;7:22.

[11] Vandersmissen B, Oramas J. On the coherency of quantitative evaluation of visual explanations. Comp Vis Im Und Jan. 2024;241:103934.

[12] Ayano YM, Schwenker F, Dufra BD, Debelee TG. Interpretable machine learning techniques in ecg-based heart disease classification. Diagn Basel Dec. 2022;13(1):111.

[13] Goettling M, Hammer A, Malberg H, Schmidt M. xecgarch: a trustworthy deep learning architecture for interpretable ecg analysis considering short-term and long-term features. Sci Rep Jun. 2024;14:13122.

Address for correspondence:

Myrte Barthels
Corda Campus, Kempische Steenweg 303/27, 3500 Hasselt
myrte.barthels@uhasselt.be