

RESEARCH ARTICLE

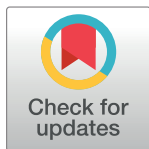
Using explainable AI to investigate electrocardiogram changes during healthy aging—From expert features to raw signals

Gabriel Ott¹ , Yannik Schaubelt¹ , Juan Miguel Lopez Alcaraz¹ , Wilhelm Haverkamp² , Nils Strodthoff¹ *

1 Carl von Ossietzky Universität Oldenburg, Oldenburg, Germany, **2** Charité Universitätsmedizin, Berlin, Germany

 These authors contributed equally to this work.

* nils.strodthoff@uol.de



OPEN ACCESS

Citation: Ott G, Schaubelt Y, Lopez Alcaraz JM, Haverkamp W, Strodthoff N (2024) Using explainable AI to investigate electrocardiogram changes during healthy aging—From expert features to raw signals. PLoS ONE 19(4): e0302024. <https://doi.org/10.1371/journal.pone.0302024>

Editor: Agnese Sbröllini, Polytechnic University of Marche: Università Politecnica delle Marche, ITALY

Received: December 18, 2023

Accepted: March 26, 2024

Published: April 11, 2024

Copyright: © 2024 Ott et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: Dataset is available from <https://physionet.org/content/autonomic-aging-cardiovascular/1.0.0/> All preprocessing files are available from <https://github.com/AI4HealthUOL/ECG-aging>.

Funding: The author(s) received no specific funding for this work.

Competing interests: The authors have declared that no competing interests exist.

Abstract

Cardiovascular diseases remain the leading global cause of mortality. Age is an important covariate whose effect is most easily investigated in a healthy cohort to properly distinguish the former from disease-related changes. Traditionally, most of such insights have been drawn from the analysis of electrocardiogram (ECG) feature changes in individuals as they age. However, these features, while informative, may potentially obscure underlying data relationships. In this paper we present the following contributions: (1) We employ a deep-learning model and a tree-based model to analyze ECG data from a robust dataset of healthy individuals across varying ages in both raw signals and ECG feature format. (2) We use explainable AI methods to identify the most discriminative ECG features across age groups. (3) Our analysis with tree-based classifiers reveals age-related declines in inferred breathing rates and identifies notably high SDANN values as indicative of elderly individuals, distinguishing them from younger adults. (4) Furthermore, the deep-learning model underscores the pivotal role of the P-wave in age predictions across all age groups, suggesting potential changes in the distribution of different P-wave types with age. These findings shed new light on age-related ECG changes, offering insights that transcend traditional feature-based approaches.

1 Introduction

Characterizing healthy aging through ECG changes

Cardiovascular diseases continue to represent the leading cause of mortality worldwide [1]. Analyzing the effect of healthy aging on the cardiovascular system enables to distinguish between an old but healthy and a younger but cardiovascular-constrained heart. This is especially difficult as there exists a discrepancy between biological age and cardiovascular age [2]. Thus, knowing what changes in the heart during healthy aging can help to avoid deaths since it enables early treatment through premature detection of cardiovascular diseases. These changes are most commonly assessed through changes in electrocardiogram (ECG) features of healthy people with age.

Shortcomings of prior work

However, by working on the feature level, relationships in the ECG that are not covered by any features may be excluded from the analysis. Previous research suggests that deep-learning models can outperform feature-based classifiers on age prediction with ECG data [3]. Studies like [4, 5] also used deep-learning models to infer the age from ECG data but they did not restrict their dataset to healthy people only. Moreover, when deep-learning models were used to do age prediction on ECG data of healthy subjects by [6], it was not analyzed how the models managed to detect the age and thus they failed to provide information on what changes in the heart during aging. Furthermore, [7–9] focused on analyzing what changes in the heart with age, but did not make use of deep-learning models.

Research questions

In this work, we aim to use techniques from explainable AI (XAI), to identify both features from feature-based age classifiers as well as ECG segments from deep learning models operating on raw data that are most important to discriminate between different age groups within a collective of healthy people. To this end, we address the following research questions: (1) Do both approaches produce insights that correspond to literature results? (2) Can this approach be used to discover new ECG feature correlations in the context of healthy aging across diverse groups?

Main contributions and findings

We analyze a dataset [10, 11] of 1,120 ECG recordings of healthy people with varying ages using two different models: an XResNet50 and an eXtreme Gradient Boosting (XGBoost) model. The XResNet50 operates on raw ECG data, while the XGBoost model uses long-range and short-range ECG features as input. Both models were trained to predict the age of a healthy person from their 1-lead-ECG and achieve a competitive performance (macro-AUCs of 0.73 and 0.77, respectively). After training, both models were then investigated with explainable AI methods where the XGBoost-model was analyzed with SHAP to find the most important ECG features for classifying the age. On the other hand, for the XResNet50 heartbeat-based saliency maps were superimposed to show the areas of interest for the age prediction task.

To summarize, our contributions and findings in this study are: (1) Our approach reaches competitive performance even though it leverages very different feature sets. (2) The XGBoost model mainly leverages long-range features. The inferred breathing rate from the ECG declines with age for healthy people and that very high SDANN5 (average standard deviation of normal-to-normal RR-intervals within a 5-minute interval) values are more likely to be from a person aged 50 or more than from a person aged 34 or less, even though SDANN values generally decline with age. (3) By construction, the XResNet is only able to leverage short-range features. Given the XResNet insights, we show that the XResNet50 exploits relationships in the P-wave with age, presumably indicating that the distribution of different types of P-wave changes with age.

2 Materials and methods

2.2 Background

ECG analysis. ECG analysis is of great importance in understanding the impact of aging on heart health and mortality. As highlighted by [12], accelerated heart-aging, as indicated by ECG-age, is associated with a significant increase in all-cause mortality, underscoring the

crucial role of ECG as a biomarker for cardiovascular risk. Investigating how the ECG of a healthy heart changes with age is therefore essential for mortality prediction. Furthermore, [9] examined ECG features to assess changes in the autonomic nervous system across various age groups, revealing clear age-related trends. While their study demonstrated such trends, it utilized a limited set of ECG features. For a comprehensive understanding of ECG analysis and its diverse applications, including critical steps like preprocessing, feature extraction, selection, transformation, and classification, [13] provides an informative survey that encompasses the breadth of this field.

Age prediction from ECGs. [6] used deep learning techniques on the automatic aging dataset [10] for age prediction. However, their approach involved reducing the original 15 age classes to only 4, which improved model performance but limited the potential for explainability insights. Meanwhile, [4] successfully predicted the age of individual subjects with a notable average error of 7 years. It is worth noting that their dataset included individuals with various health conditions, raising the possibility of age inference from age-related diseases. A similar result was achieved on a public dataset in [5]. [3] introduced a promising predictive approach comparing deep learning models working on raw ECG data and tree-based classifiers using ECG features. These models were trained on a substantial dataset of over 2.3 million 12-lead ECGs for diverse tasks. However, the study did not explore the explainability of their models, which remains an important aspect for further investigation in age prediction from ECG data. Our research addresses these gaps by considering a finer granularity of age groups, allowing for more detailed insights from explainable methods. Additionally, by focusing solely on data from healthy individuals, we ensure cleaner data and more reliable insights and hence less confounding factors, thus contributing to a more comprehensive understanding of age prediction from ECGs.

ECG explainability. A recent review [14] highlights the potential of using techniques of explainable AI (XAI) to uncover mechanisms underlying age prediction models, an approach that resonates very well with the approach taken in this work. In the domain of ECG analysis, the significance of interpretability has gained prominence in recent studies, as nicely reviewed in recent systematic reviews [15]. Researchers have been actively integrating XAI techniques to enhance the interpretation of ECG data. However, in many cases, this crucial component gets reduced to anecdotal evidence obtained from the straightforward application of commonly used attribution methods to handpicked examples to underline the validity of the proposed algorithm. On the contrary, two recent dedicated works on interpretability in the ECG domain [16, 17] highlight the methodology of aggregated attributions across patients or entire patient populations, a technique that is also supposed to be used in this study.

2.2 Dataset and data preparation

Dataset. The Autonomic Aging dataset [10, 11] aims to quantify changes in cardiovascular autonomic function during healthy aging. It contains ECG recordings of 1,120 healthy-control subjects sampled at 1,000 Hz in a resting state under controlled measurement conditions. Nevertheless, for the purpose of this study, we considered only ECGs where the age information is given, leaving 1,095 patients. The patients' ages range from 18 to 92 years, and the recordings span from 8 to 35 minutes, with a mean of 19 minutes. Two different devices were used to measure the ECGs, a 1-lead and a 2-lead ECG recorder. Therefore, we used only the matching lead (II) from both devices. The gender distribution within the dataset shows a slight imbalance towards female patients (675 female and 420 male patients). The dataset contains 38 ECG recordings with missing values, however, most of them with only a few time steps across the complete signal at 1000 Hz, therefore, missing values were removed without excluding

recordings. For approaches based on raw time series data, we work at a temporally down-sampled resolution of 100 Hz, which was found to be sufficient for common diagnostic tasks [18], and also removed a negligible number of missing values in the time series.

In the dataset, all subjects are diagnosed as healthy, so we sampled 3-second crops in order to capture at least one complete heartbeat. Due to the imbalanced nature of the dataset, we opted for a 60/20/20 split between training, validation, and test sets at the subject level. Importantly, we maintained the age-group distribution in each set as far as possible while simultaneously ensuring that every age group was represented in all three sets. For consistency, all models presented in this work use identical splits for their training, validation, and test data. Table 1 contains additional descriptive statistics of the processed dataset. See our source code for the dataset pre-processing steps [19] for more details.

Age-group distribution. Fig 1 shows the age distribution across the dataset in terms of 15 age groups, where the first age group contains subjects aged 18 to 19, whereas all following age groups but the last cover age intervals of 5 years. There is a clear imbalance in the age distribution, with the majority in age group 20–24 with 422 samples, followed by age group 25–29 with 105 samples. On the contrary, the last four classes represent only 39 samples or correspondingly 3.4% of the full dataset. Furthermore, it is important to mention that there are no male samples available for the age groups 75–79 years and 85–92 years. As past studies did not indicate a strong interaction effect between gender and age prediction, and dividing the dataset by gender would worsen the imbalance in the smaller age groups, we decided to ignore gender as a covariant in this study.

2.3 Models and feature sets

Overview. For the purpose of this study, we investigate two different classifiers working on two different feature sets but both predicting the subject's age: a residual neural network (XResNet50) operating on raw time series data and a tree-based model gradient boost decision tree classifier (XGBoost) operating on derived features. We carried out diverse experiments to find sufficient feature sets and preprocessing settings which could lead us to a scenario of

Table 1. Summary of the dataset composition.

	Train	Validation	Test
Subjects			
Male	251 (38.20%)	84 (38.35%)	85 (38.81%)
Female	406 (61.80%)	135 (61.65%)	134 (61.18%)
Age			
Median	25–29y	25–29y	25–29y
Quantile 25%	20–24y	20–24y	20–24y
Quantile 75%	35–39y	35–39y	35–39y
Patient ECG record in seconds			
Median	920.74	922.51	916.3
Quantile 25%	900.3	900.23	899.65
Quantile 75%	1207.7	1222.5	1204.89
Mean	1077.39	1112.9	1053.24
Std. Dev.	349.72	387.24	343.15
Min	599.7	480.5	600.2
Max	2056.7	2051.66	2168.2
Total crops	235947	81241	76886

<https://doi.org/10.1371/journal.pone.0302024.t001>

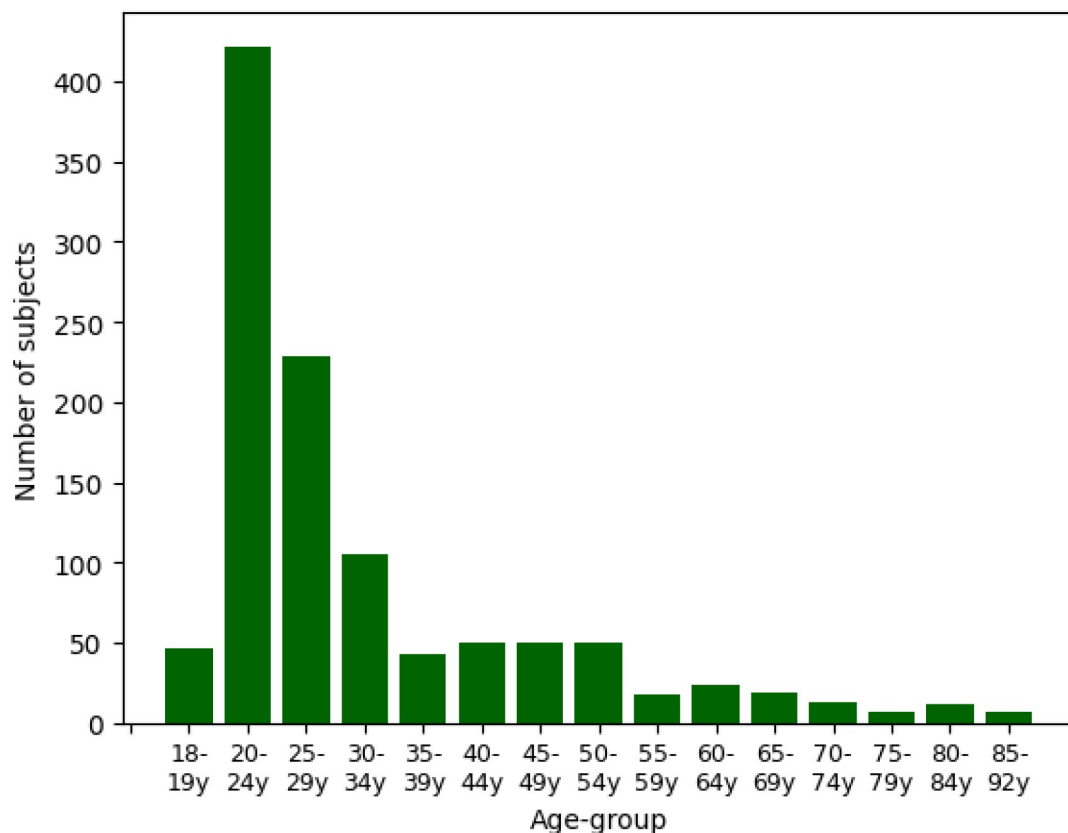


Fig 1. Age-group distribution. Age-group distribution in terms of age groups provided in the Autonomic Aging dataset [10]. The age groups span a range from 18 to 92 years, where the majority of patients are between 20 to 50 years old.

<https://doi.org/10.1371/journal.pone.0302024.g001>

better performance results. In each of the settings, we report test set scores of models selected based on held-out validation set scores. To facilitate continued research, we release the source code underlying our study [19].

2.3.1 XGBoost. For the XGBoost model, we include long-range heart-rate variability (HRV) and short-range (SR) features. The HRV features describe how the ECG signal varies over time, and it contains features such as Standard Deviation of NN Intervals (SDNN), Root Mean Square of Successive Differences (RMSSD) and low and high-frequency powers to name a few. The HRV features were extracted from the ECGs with NeuroKit2 [20]. The SR features were calculated from fiducial points and comprise features such as R-R intervals, heart rate, peak amplitudes, and waves such as Q, R and S, and P and T respectively. The SR features were extracted per heartbeat with the python HeartPy toolkit [21] in combination with NeuroKit2. To produce an age prediction for a whole ECG, the heartbeat-interval SR features were averaged over each ECG recording. This enabled us to combine the SR and HRV feature sets. As there are two different feature sets, all three combinations were tested: each alone and both sets combined.

Training. As a countermeasure against the label imbalance in the training dataset, we created artificially balanced datasets by oversampling the minority classes with a random oversampling technique. Lastly, regarding the model training, we performed a grid search to determine optimal hyperparameters based on validation set performance. After this process, the only hyperparameters where we found deviations from default values to be beneficial were `max_depth = 10`, `max_leaves = 10`, `learning_rate = 0.008`. For the explainability analysis, we

leverage SHAP values [22]. This is in line with a recent comparative study [23] where SHAP values showed good overlap with cardiologists' expert features.

Relevant features. At this point, it is worthwhile explicitly highlighting a number of ECG features that will play an important role for the later analysis:

- *SDNN and SDNN5*: SDNN represents the standard deviation of normal-to-normal RR-intervals, while SDNN5 denotes the average SDNN calculated within a 5-minute interval. Similarly, SDANN1 refers to the average SDNN computed within a 1-minute interval.
- *HRV_PAS*: This metric quantifies the percentage of NN intervals within alternating segments, where NN intervals represent the time intervals between normal R-peaks in the ECG signal.
- *Alpha-Features*: Alpha-features are derived from detrended fluctuation analysis (DFA) and provide insights into the auto-correlation between heartbeats. Specifically, alpha1 characterizes short-term correlations, while alpha2 captures long-term correlations in heart rate variability.
- *pNN*: pNN20 signifies the percentage of heartbeat intervals with more than a 20-millisecond deviation from the previous interval, while pNN50 represents the corresponding percentage for intervals with more than a 50-millisecond deviation.
- *MCVNN*: MCVNN stands for the median absolute deviation of RR intervals divided by the median of RR intervals, providing valuable information about heart rate variability.

These ECG features serve as critical components for our analysis, and understanding their definitions is essential for comprehending the subsequent sections of this paper.

2.3.2 XResNet50. The XResNet50 deep learning model works with raw time series data. We chose the XResNet50 model, which showed competitive performance with the best-performing convolutional neural networks for a range of different ECG classification tasks [5, 24]. It represents a one-dimensional adaptation of a commonly used ResNet-type convolutional neural network from computer vision [25]. Here we additionally restrict to a single input channel as appropriate for 1-lead ECG data.

Training. The XResNet50 was trained with the AdamW optimizer and weight decay [26]. We investigate two different loss functions, namely focal loss (FL) [27] and cross-entropy loss (CEL). The learning rate was set to 10^{-5} for FL and 10^{-2} for CEL and adjusted with a reduced learning rate on the plateau scheduler, which divided the learning rate by 10 if the loss did not decrease for 2 consecutive epochs. We trained on 20 epochs with early stopping after 3 consecutive epochs. Oversampling the training set produced insufficient performance results in early experiments. Thus, as a countermeasure against the unbalanced training set, class weights are applied instead. The class weights were set to the inverse of each age group's number of occurrences in the training set. We investigate different scenarios, firstly by using two different loss functions and, secondly, by applying training class weights.

We trained the models on crops of 3s length and aggregate predictions from multiple crops using mean output predictions to obtain sample-level predictions, see [18] for a detailed analysis of the benefits of this procedure. Note that using 3s-crops limits the XResNet50 to detect short-range patterns. We leverage the methodology proposed in [17] to compute beat-aligned attribution maps over entire patient subgroups. In particular, we use saliency maps as attribution maps as saliency was the only attribution method that satisfied the sanity checks proposed in [17].

2.4 Performance metric

For comparability with earlier works, we report accuracy as a performance metric but stress the severe shortcomings of accuracy in the presence of severe class imbalance as is the case here. As

the main performance metric, we report the macro-averaged (over age groups) area under the receiver operating curve (macro-AUC), which is less affected by class imbalance and operates on output probabilities rather than dichotomized outputs. To assess the uncertainty of our predictions due to the finite size of the test set, we resort to bootstrapping on the test set. We report 2.5 and 97.5 percentiles of the test set scores, i.e. 90% confidence intervals for the test set scores. We indicate these within brackets behind the point estimate for the score such as 0.05(0.025,0.975).

3 Results

3.1 Predictive performance results

3.1.1 XGBoost. In Table 2, we present the performance evaluation of the XGBoost model operating on different feature sets, including SR, HRV, and a combined feature set HRV+SR, trained on both balanced (oversampled) and unbalanced (original) datasets. The results indicate a nuanced performance pattern, with the balanced configuration generally exhibiting slightly superior performance compared to the unbalanced setting. Notably, the model operating on SR features performed worse with an AUC of 0.70. Conversely, the model incorporating both SR and HRV features and trained on a balanced dataset demonstrated the highest efficacy, achieving an AUC score of 0.77 on the test set. The performance gain over the model leveraging on HRV features confirms that the combined model actually exploits both short-range as well as long-range features.

To set these results into perspective, we also show a direct comparison between our feature-based XGBoost model and a previously introduced feature-based approach [6]. For comparability, we follow their approach and consolidate the original 15 age groups into 4 broader categories. Both models demonstrate closely aligned accuracy scores, with our XGBoost model achieving 0.684 (95% CI: 0.62–0.74) and the prior feature-based model at 0.688 (95% CI: 0.64–0.73). The similarity in performance with almost identical point estimates and largely overlapping confidence intervals underscores the parity between our XGBoost model and the established feature-based approach. This reinforces the reliability of our findings and underscores the suitability of our model for age group classification tasks, laying the foundation for further explainability investigations.

3.1.2 XResNet. Table 3 presents the performance evaluation for different XResNet50 configurations. Notably, the experimental findings underscore the superiority of focal loss over cross-entropy loss as the preferred choice for this task with a severely imbalanced label distribution. Furthermore, irrespective of the loss function employed, it is evident that the model attains significantly enhanced performance levels when trained on the unbalanced training

Table 2. XGBoost model macro-AUC on the test set for different configurations.

	Balanced	Unbalanced
HRV	0.74 (0.69, 0.79)	0.73 (0.68, 0.77)
SR	0.70 (0.63, 0.75)	0.70 (0.65, 0.74)
HRV+SR	0.77(0.72, 0.80)	0.72 (0.68, 0.77)

Table 3. XResNet50 performance on different configurations.

	Balanced	Unbalanced
FL	0.71 (0.67, 0.75)	0.74(0.70, 0.77)
CEL	0.62 (0.57, 0.67)	0.65 (0.60, 0.70)

<https://doi.org/10.1371/journal.pone.0302024.t003>

dataset. The most noteworthy configuration emerges with focal loss applied to the unbalanced training dataset, achieving a commendable macro AUC score of 0.74. It is worth stressing that the training and validation happened on crop level but the testing on subject level. By averaging the prediction of all crops belonging to a patient, a subject-level prediction was formed.

3.1.3 Comparative assessment. When comparing the results from both models it is interesting to see that both reach a comparable performance despite fundamentally different input representations and model architectures. The most direct comparison is between the XGBoost model operating on SR (short-range) and the XResNet model, which by construction only leverages short-range features as well. It reveals a slight advantage on the side of the XResNet model, which is in line with the original hypothesis that the raw waveform contains additional discriminative information that is not covered in conventionally considered short-range ECG features. The long-range information contained in the HRV features is somewhat complementary to this information as the increase in the predictive performance of the HRV+SR model compared to the SR model shows. It remains an interesting question for future research if such long-range interactions could also be exploited using models operating on raw time series data with appropriate model architectures, see [18] for first steps in this direction.

Additionally, Fig 2 displays sample-level AUC scores for both models across diverse age groups, revealing a general performance improvement with increasing age, albeit with a notable exception in the 75-79-year age group for the XResNet model and lower scores observed within the 24 to 44-year age range. These findings provide valuable insights into the nuanced age-related patterns discerned by both models.

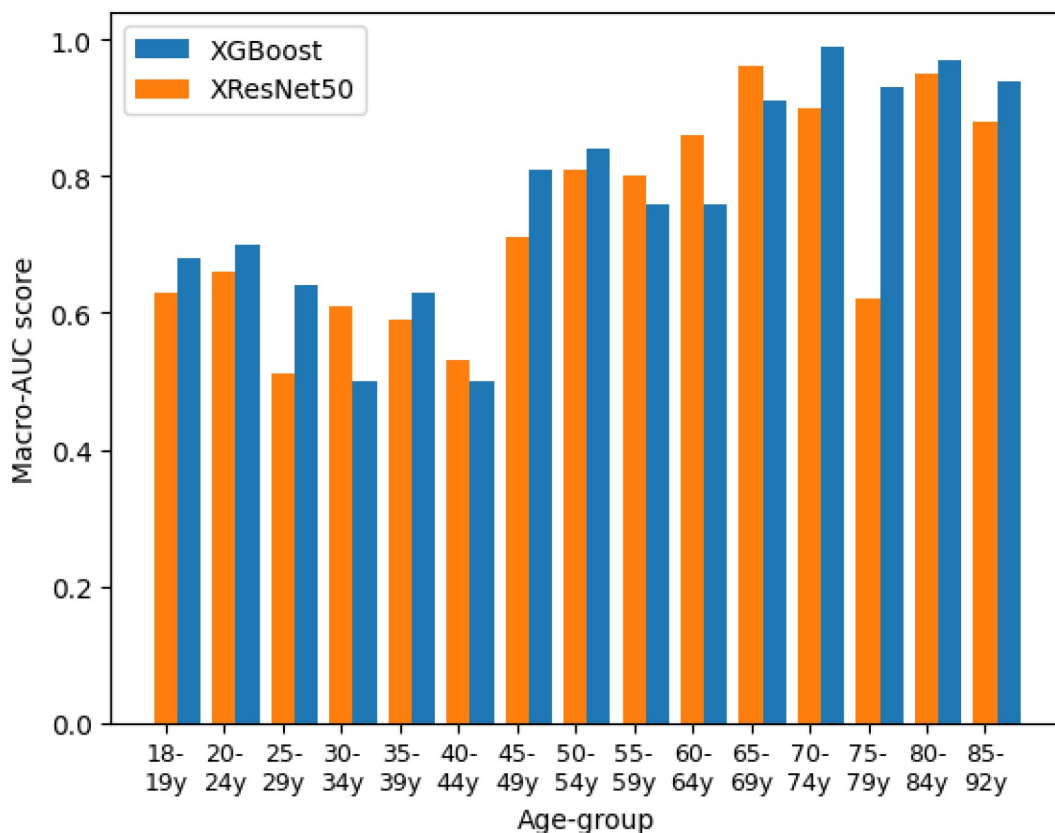


Fig 2. Predictive performance. Predictive performance results of the models per age group in terms of AUC on the test set, where the yellow (left) age group represents the XGBoost and the right (blue) the XResNet.

<https://doi.org/10.1371/journal.pone.0302024.g002>

3.2 Explainability results

3.2.1 XGBoost. SHAP feature relevances. In this study, we classify the features into two categories: those with the prefix HRV- denote long-range HRV features, while those with the prefix SR- represent short-range features. Subsequently, we analyze the top 10 influential features for each age group, see Fig 3.

Important features with consistent age trends. Observations reveal that certain features are recurrent across multiple age groups, displaying a consistent trend with respect to age. Specifically, HRV_SDANN5, HRV_PAS, P-wave amplitude (p_mV) and alpha-fluctuation values consistently increase with age where lower values of these features are indicative of younger individuals, whereas higher values are associated with elderly individuals. Conversely, certain other features exhibit a contrasting trend. For instance, pNN20, MCVNN, and breathing rate in conjunction with breathing signal exhibit a decline with advancing age. High values of these features correspond to younger individuals, while lower values are characteristic of older individuals. Note that in the case of breathing rate and breathing signal, these features are derived from the ECG and serve as estimates of respiratory activity.

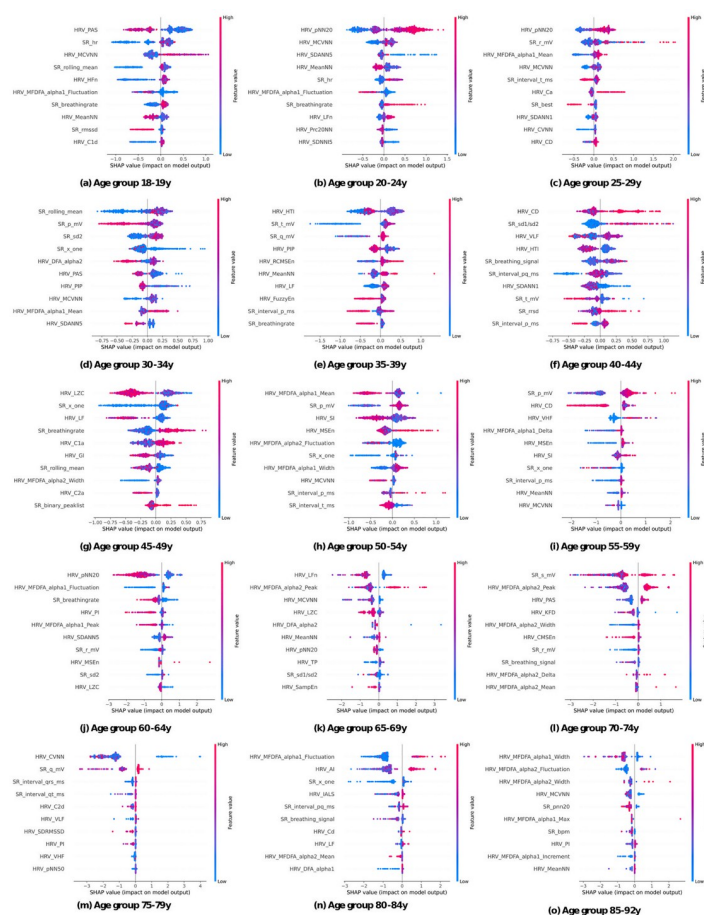


Fig 3. SHAP values across age groups. The 10 most important features for classifying the age groups are depicted in these subplots. In each subplot, the features are arranged in descending order of importance, emphasizing their significance in age group classification. The color scheme, with blue dots representing low feature values and red dots denoting high feature values, provides a visual representation of the feature's influence across different age groups. As you move from left to right and from top to bottom, you explore the SHAP values for all age groups.

<https://doi.org/10.1371/journal.pone.0302024.g003>

Consistency with literature results. It is noteworthy that our findings align with existing research in several aspects. Specifically, the observed trends in pNN, alpha-mean, HRV-PAS, P-wave amplitude and alpha-fluctuation are consistent with previous studies. For instance, the decrease in pNN50 with age among healthy subjects, as well as the discriminative power of pNN50 and pNN20 in age separation, has been noted by [28]. Similarly, the steady increase in alpha-values with age among healthy subjects, as well as rising alpha-fluctuations, has been reported [29, 30], albeit without specific reference to alpha-mean. Furthermore, the upward trajectory of HRV-PAS with age as observed in our model, is consistent with the findings of [31]. Similarly, the pattern of P-wave amplitude rising up to age 60 before declining, as observed in our study, concurs with the research of [32]. In summary, our XGBoost model's conclusions are in accordance with existing research, reinforcing the notion that certain physiological features exhibit consistent age-related trends, which can be valuable in understanding the physiological changes associated with aging.

New insights: Breathing rate and SDANN5. At this stage we have presented parts of our findings that align with previous research, however, we further provide insights into ECG and healthy aging, specifically for breathing rate and SDANN trends.

According to [33] breathing rate and age are hardly correlated at all. However, the 2.5-97.5 percentile of the breathing rate increases with age according to [34], meaning that lower- and higher breathing rates become more common with age. Furthermore, [35] suggests that there are variations in respiratory dynamics, particularly in response to metronome breathing such as the increase in high frequency at different postural changes, especially in young subjects. In contrast to this work, these studies were not limited to healthy subjects only. Since all subjects in this work are healthy this means that the breathing rate decreases with age for healthy subjects. This finding is also plausible when considering that all subjects were in a resting state during recording. Because of the general decline in body activity with age less energy and thus oxygen is required to run body activities in a resting state. Assuming that the lungs and heart are healthy a lower breathing rate is therefore plausible for healthy aging.

Our model reveals interesting insights regarding the SDANN5 feature. Contrary to established research showing a general decline in SDANN with age, the explainability analysis suggests that high SDANN5-values contribute positively towards the age prediction of older individuals. For instance, it associates lower SDANN5 values with those aged 20-34 and higher values with those aged 60-64. A similar study using the same dataset [6] shows that while the mean SDANN does decline with age, there are significant variations in SDANN values among age groups, which lead to very high SDANN values being more probable for subjects older than 50 compared to subjects younger than 30. In summary, our XGBoost model uncovers an unexpected relationship between age and SDANN5, challenging the conventional wisdom of decreasing SDANN values with age. Notably, this effect is primarily observed in specific age groups beyond age 60.

3.2.2 XResNet. *Beat-level descriptive analysis.* At first, we explore superimposed mean heartbeats for all age groups in Fig 4 as a plausibility test and to compare with literature statements. The amplitude of the T-wave decreases with age and shifts to the right, indicating an overall longer cardiac cycle, meaning a slower heart rate. Furthermore, the T- and P-wave intervals shorten with age; moreover, the absolute magnitude of the S-peak, Q-peak, and P-wave appears to diminish with age as well, which is in accordance with [36, 37]. It is noteworthy that the amplitude of the R-peak shows no conclusive trend with age.

Aggregated saliency maps: Methodology. Since ECGs even in the same age group have slightly different heart rates and are generally not aligned, the crop-level-saliency maps cannot simply be laid on top of each other. Following [17], the crops of each subject with their saliency maps were split into individual heartbeats and averaged from 30 milliseconds before to 50

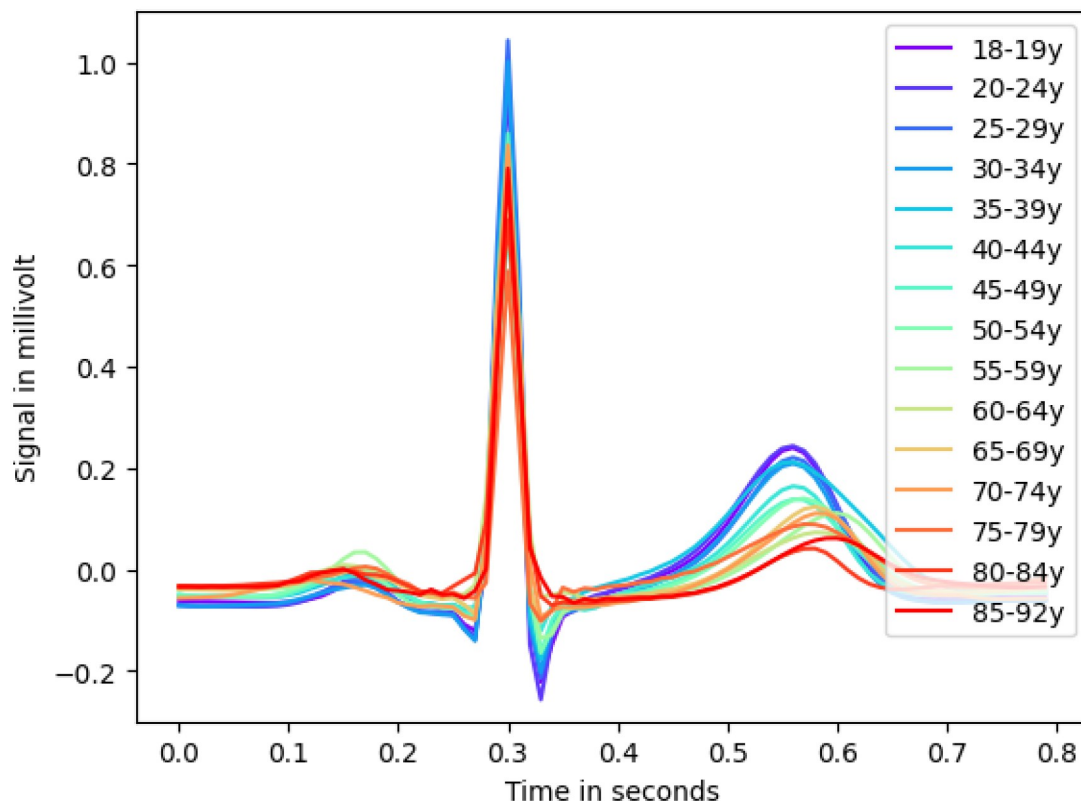


Fig 4. Aggregated mean heartbeat. Aggregated mean heartbeat for all age groups showcases ECG feature trends across age groups.

<https://doi.org/10.1371/journal.pone.0302024.g004>

milliseconds after the R-peak. Then, these medium heartbeats were again averaged for each age group, resulting in one aggregated heartbeat per age group as shown in Fig 5. To reveal the patterns exploited by the model most clearly, we used the training set to produce the aggregated attribution maps. We also mark the most salient data points (marked in red) to identify patterns across age groups as described below.

Aggregated saliency maps: Results. The XResNet model consistently demonstrates a prediction for the entire P-wave as individuals age. It specifically focuses on the offset, with some onset in early age groups. These variations may reflect different P-wave types. Prior studies found that, the distribution of which undergoes significant changes with age. [32]. Furthermore, research by [38–40] has elucidated age-related disparities in various aspects of the P-wave, including its duration. Consequently, it is plausible that the Deep Learning model distinguishes age groups based on distinct P-wave parts and their respective distributions, underlining the complexity of its age classification methodology. The application of more sophisticated methods, for example from the domain of concept-based XAI such as [41], would be the logical next step to uncover these changes. Apart from the P-wave, the model frequently focuses on the Q-peak and S-peak while showing limited relevance to the R-peak and the peak of the T-wave. The TP segment receives moderate attention, indicating its importance in age-related classification.

3.2.3 Comparative assessment. Table 4 presents key insights into age-related trends of ECG features that were derived from applying XAI on the XGBoost and XResNet50 models.

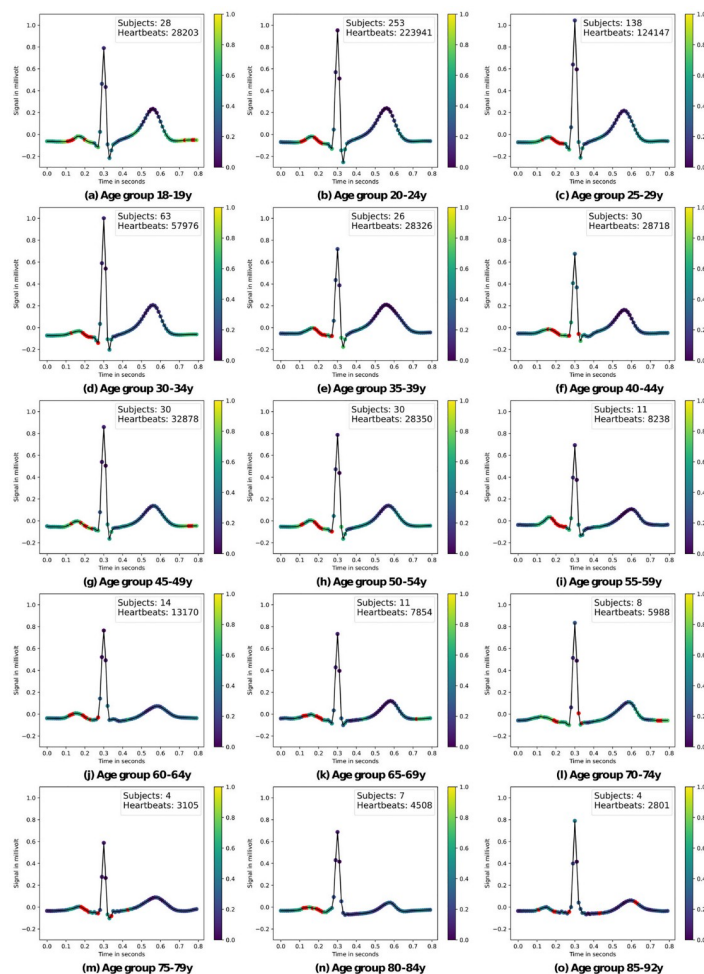


Fig 5. Saliency maps across age groups. Within each subplot, as you progress from left to right and from top to bottom, you navigate through the beat-level ECG saliency maps, shedding light on the key ECG features that contribute to age group differentiation according to the colour map scheme. 'Subjects' and 'Heartbeats' state the number of subjects and heartbeats used to create these plots. The eight highest gradients are marked in red.

<https://doi.org/10.1371/journal.pone.0302024.g005>

In the XGBoost model, age is associated with a decrease in pNN20, MCVNN, and breathing rate, along with an increase in alpha-fluctuations, P-wave amplitude, and PAS. Additionally, SDANN5 values rise with age. In contrast, the XResNet50 exhibits distinct focus areas with age: given our criterion over the eight most important saliency time steps, it emphasizes P-wave features (18.33% for onsets and 53.33% for offsets), frequently places relevance on the Q-wave peak (8.33%), shows little relevance on the R-peak, sometimes focuses on the S-peak (4.16%), attributes minimal relevance to the T-wave (3.33%), and frequently assesses the TP-interval (12.5%). The latter might be related to differences in the heart rate, which are difficult to analyze by means of saliency maps. Nevertheless, the observed trends offer valuable insights into age group differentiation in the two models' ECG interpretations.

3.2.4 Data imbalance and research focus. While the 'autonomic aging' dataset [10] used in this work stands out as one of the largest datasets of its kind, it is important to acknowledge its inherent imbalance, notably the scarcity of samples from individuals aged 70 or older. Deep learning models, with their appetite for ample training data, face a particular challenge in such

Table 4. Clinical observations and feature trends in age-related explainability. Percentages refer to relative number of age groups where a high-saliency timestep (red in Fig 5) occurred in the corresponding segment.

XGBoost feature trends	XResNet50 observations
pNN20 ↓	Strong focus on P-wave offsets (53.33%), with some P-onsets (18.33%) in early age groups.
Alpha-fluctuations ↑	Focus on Q-peaks (8.33%), especially in middle-aged groups.
MCVNN ↓	R-peak is largely ignored by the model.
P-wave amplitude ↑	S-peak (4.16%) is important for age groups 35–44y and 70–79y.
PAS ↑	Hardly any focus on T-wave (3.33%).
breathing rate and signal ↓	Frequent focus on TP-segment (12.5%).
SDANN5 ↑	

<https://doi.org/10.1371/journal.pone.0302024.t004>

scenarios. Addressing the imbalance by consolidating the underrepresented age groups might seem like a logical step, however, leads to a less nuanced prediction model. We have deliberately chosen not to merge age groups above 70 years as our primary focus centers on understanding the nuances of a healthily aging heart. In this context, we find that the uniqueness of our dataset, even with its imbalances, continues to yield more insightful results that better align with our research objectives.

4 Conclusion

In this study, we investigated age-related cardiovascular changes of a healthy population. Leveraging ECG data from 1,095 healthy subjects, we developed two different models that work across diverse data modalities and used feature attribution methods to study their behavior: an XGBoost model analyzing short- and long-range ECG features as well as an XResNet50 model processing raw ECG data. Our experiments suggest that the feature-based model achieves better predictive performance in comparison with raw ECG data, which is comparable with literature performance. The findings from the feature-based model indicate increasing heart irregularity and reduced flexibility with age, which aligns with prior research. It also revealed a decline in inferred breathing rate with age and the significance of high SDANN values in older individuals. Notably, the deep-learning model identified the P-wave as the most important segment across all age groups. Our findings provide complementary insights into age-related ECG changes, whose identification is crucial for the early detection of cardiovascular diseases. To promote further exploration in this area of study, we release the source code underlying our study [19].

Author Contributions

Formal analysis: Juan Miguel Lopez Alcaraz.

Investigation: Gabriel Ott, Yannik Schaubelt, Juan Miguel Lopez Alcaraz.

Supervision: Juan Miguel Lopez Alcaraz, Nils Strodthoff.

Validation: Wilhelm Haverkamp.

Visualization: Gabriel Ott, Yannik Schaubelt.

References

1. Vaduganathan M., Mensah G. A., Turco J. V., Fuster V., and Roth G. A., "The global burden of cardiovascular diseases and risk," *Journal of the American College of Cardiology*, vol. 80, no. 25, pp. 2361–2371, 2022. <https://doi.org/10.1016/j.jacc.2022.11.005>

2. Pavanello S., Campisi M., Fabozzo A., Cibin G., Tarzia V., Toscano G., et al. "The biological age of the heart is consistently younger than chronological age," *Scientific Reports*, vol. 10, no. 1, Jul. 2020.
3. Zvuloni E., Read J., Ribeiro A. H., Ribeiro A. L. P., and Behar J. A., "On merging feature engineering and deep learning for diagnosis, risk prediction and age estimation based on the 12-lead ECG," *IEEE Transactions on Biomedical Engineering*, vol. 70, no. 7, pp. 2227–2236, Jul. 2023.
4. Attia Z. I., Friedman P. A., Noseworthy P. A., Lopez-Jimenez F., Ladewig D. J., Satam G., et al. "Age and sex estimation using artificial intelligence from standard 12-lead ECGs," *Circulation: Arrhythmia and Electro-physiology*, vol. 12, no. 9, Sep. 2019. <https://doi.org/10.1161/CIRCEP.119.007284> PMID: 31450977
5. Strodthoff N., Wagner P., Schaeffter T., and Samek W., "Deep learning for ecg analysis: Benchmarks and insights from ptb-xl," *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 5, pp. 1519–1528, 2020.
6. Lee K. H. and Byun S., "Age prediction in healthy subjects using RR intervals and heart rate variability: A pilot study based on deep learning," *Applied Sciences*, vol. 13, no. 5, p. 2932, Feb. 2023.
7. Schumann A., Gaser C., Sabeghi R., Schulze P. C., Festag S., Spreckelsen C., et al. "Using machine learning to estimate the calendar age based on autonomic cardiovascular function," *Frontiers in Aging Neuroscience*, vol. 14, Jan. 2023.
8. Garavaglia L., Gulich D., Defeo M. M., Mailland J. T., and Iurzun I. M., "The effect of age on the heart rate variability of healthy subjects," *PLOS ONE*, vol. 16, no. 10, p. e0255894, Oct. 2021. <https://doi.org/10.1371/journal.pone.0255894>
9. S. M. Z. S. Bukhari and I. Akhtar, "HRV dynamical analysis of cardiac autonomic activity of healthy subjects with age," in *2023 International Conference on Robotics and Automation in Industry (ICRAI)*. IEEE, Mar. 2023.
10. Schumann A. and Bär K.-J., "Autonomic aging—a dataset to quantify changes of cardiovascular autonomic function during healthy aging," *Scientific Data*, vol. 9, no. 1, p. 95, Mar. 2022. <https://doi.org/10.1038/s41597-022-01202-y>
11. Goldberger A. L., Amaral L. A. N., Glass L., Hausdorff J. M., Ivanov P. C., Mark R. G., et al. "Physio-Bank, PhysioToolkit, and PhysioNet," *Circulation*, vol. 101, no. 23, pp. e215–e220, 2000.
12. L. C. Brant, A. H. Ribeiro, M. M. Pinto-Filho, J. Kornej, S. R. Preis, J. L. Fetterman, et al. "Association between electrocardiographic age and cardiovascular events in community settings: The framingham heart study," *Circulation: Cardiovascular Quality and Outcomes*, Jun. 2023.
13. Merdjanovska E. and Rashkovska A., "Comprehensive survey of computational ECG analysis: Data-bases, methods and applications," *Expert Systems with Applications*, vol. 203, p. 117206, Oct. 2022. <https://doi.org/10.1016/j.eswa.2022.117206>
14. A. Kalyakulina, I. Yusipov, M. G. Bacalini, A. Moskalev, C. Franceschi, and M. Ivanchenko, "explainable artificial intelligence (xai) in aging clock models," 2023.
15. Ayano Y. M., Schwenker F., Dufera B. D., and Debelee T. G., "Interpretable machine learning techniques in ECG-based heart disease classification: A systematic review," *Diagnostics*, vol. 13, no. 1, p. 111, Dec. 2022.
16. Bender T., Beinecke J. M., Krefting D., Müller C., Dathe H., Seidler T., 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*, pp. 1–12, 2023.
17. P. Wagner, T. Mehari, W. Haverkamp, and N. Strodthoff, "Explaining deep learning for ecg analysis: Building blocks for auditing and knowledge discovery," *arXiv preprint arXiv:2305.17043*, 2023.
18. Mehari T. and Strodthoff N., "Towards quantitative precision for ECG analysis: Leveraging state space models, self-supervision and patient metadata," *IEEE Journal of Biomedical and Health Informatics*, pp. 1–9, 2023.
19. G. Ott, Y. Schaubelt, J. M. Lopez Alcaraz, W. Haverkamp, and N. Strodthoff, "ECG-aging," Oct. 2023. [Online]. Available: <https://github.com/AI4HealthUOL/ECG-aging>.
20. Makowski D., Pham T., Lau Z. J., Brammer J. C., Lespinass F., Pham H., et al. "NeuroKit2: A python toolbox for neurophysiological signal processing," *Behavior Research Methods*, vol. 53, no. 4, pp. 1689–1696, Feb. 2021. <https://doi.org/10.3758/s13428-020-01516-y>
21. Gent P. V., Farah H., Nes N. V., and Arem B. V., "Analysing noisy driver physiology real-time using off-the-shelf sensors: Heart rate analysis software from the taking the fast lane project," *Journal of Open Research Software*, vol. 7, no. 1, p. 32, Oct. 2019. <https://doi.org/10.5334/jors.241>
22. Lundberg S. M., Erion G., Chen H., DeGrave A., Prutkin J. M., Nair B., et al. "From local explanations to global understanding with explainable ai for trees," *Nature machine intelligence*, vol. 2, no. 1, pp. 56–67, 2020. <https://doi.org/10.1038/s42256-019-0138-9> PMID: 32607472
23. T. Mehari, A. Sundar, A. Bosnjakovic, P. Harris, S. E. Williams, A. Loewe, et al. "Ecg feature importance rankings: Cardiologists vs. algorithms," *arXiv preprint arXiv:2304.02577*, 2023.

24. Mehari T. and Strodthoff N., "Self-supervised representation learning from 12-lead ECG data," *Computers in Biology and Medicine*, vol. 141, p. 105114, Feb. 2022. <https://doi.org/10.1016/j.combiomed.2021.105114>
25. T. He, Z. Zhang, H. Zhang, Z. Zhang, J. Xie, and M. Li, "Bag of tricks for image classification with convolutional neural networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 558–567.
26. I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *International Conference on Learning Representations*, 2018.
27. T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–2988.
28. Mietus J. E., "The pNNx files: re-examining a widely used heart rate variability measure," *Heart*, vol. 88, no. 4, pp. 378–380, Oct. 2002. <https://doi.org/10.1136/heart.88.4.378> PMID: 12231596
29. O. Barquero-Perez, J. M. de Sa, J. L. Rojo-Alvarez, and R. Goya-Esteban, "Changes in detrended fluctuation indices with aging in healthy and congestive heart failure subjects," in *2008 Computers in Cardiology*. IEEE, Sep. 2008.
30. Iyengar N., Peng C. K., Morin R., Goldberger A. L., and Lipsitz L. A., "Age-related alterations in the fractal scaling of cardiac interbeat interval dynamics," *American Journal of Physiology-Regulatory, Integrative and Comparative Physiology*, vol. 271, no. 4, pp. R1078–R1084, Oct. 1996.
31. Costa M. D., Davis R. B., and Goldberger A. L., "Heart rate fragmentation: A new approach to the analysis of cardiac interbeat interval dynamics," *Frontiers in Physiology*, vol. 8, May 2017.
32. Havmoller R., Carlson J., Holmqvist F., Herreros A., Meurling C. J., Olsson B. et al. "Age-related changes in p wave morphology in healthy subjects," *BMC Cardiovascular Disorders*, vol. 7, no. 1, Jul. 2007.
33. Takayama A., Nagamine T., and Kotani K., "Aging is independently associated with an increasing normal respiratory rate among an older adult population in a clinical setting: A cross-sectional study," *Geriatrics and Gerontology International*, vol. 19, no. 11, pp. 1179–1183, Oct. 2019. <https://doi.org/10.1111/ggi.13788>
34. Rodríguez-Molinero A., Narvaiza L., Ruiz J., and Gálvez-Barrón C., "Normal respiratory rate and peripheral blood oxygen saturation in the elderly population," *Journal of the American Geriatrics Society*, vol. 61, no. 12, pp. 2238–2240, Dec. 2013. <https://doi.org/10.1111/jgs.12580>
35. Schwartz J. B., Gibb W. J., and Tran T., "Aging Effects on Heart Rate Variation," *Journal of Gerontology*, vol. 46, no. 3, pp. M99–M106, May 1991. <https://doi.org/10.1093/geronj/46.3.M99> PMID: 2030273
36. Zareba K. M., Truong V. T., Mazur W., Smart S. M., Xia X., Couderc J.-P., et al. "T-wave and its association with myocardial fibrosis on cardiovascular magnetic resonance examination," *Annals of Noninvasive Electrocardiology*, vol. 26, no. 2, Dec. 2020. <https://doi.org/10.1111/anec.12819> PMID: 33336876
37. Bocchi F., Marques-Vidal P., Pruvot E., Waeber G., Vollenweider P., and Gachoud D., "Clinical and biological determinants of p-wave duration: cross-sectional data from the population-based CoLausPsy-CoLaus study," *BMJ Open*, vol. 10, no. 11, p. e038828, Nov. 2020. <https://doi.org/10.1136/bmjopen-2020-038828>
38. Giovanardi P., Vernia C., Tincani E., Giberti C., Silipo F., and Fabbo A., "Combined effects of age and comorbidities on electrocardiographic parameters in a large non-selected population," *Journal of Clinical Medicine*, vol. 11, no. 13, 2022.
39. Turhan H., Yetkin E., Sahin O., Yasar A. S., Senen K., Atak R., et al. "Comparison of p-wave duration and dispersion in patients aged ≥ 65 years with those aged ≤ 45 years," *Journal of Electrocardiology*, vol. 36, no. 4, pp. 321–326, Oct. 2003. <https://doi.org/10.1016/j.jelectrocard.2003.08.002>
40. Lindow T., Palencia-Lamela I., Schlegel T. T., and Ugander M., "Heart age estimated using explainable advanced electrocardiography," *Scientific Reports*, vol. 12, no. 1, p. 9840, 2022. <https://doi.org/10.1038/s41598-022-13912-9>
41. J. Vielhaben, S. Blücher, and N. Strodthoff, "Multi-dimensional concept discovery (MCD): A unifying framework with completeness guarantees," *Transactions on Machine Learning Research*, 2023. [Online]. Available: <https://openreview.net/forum?id=KxBQPz7HKh>.