# Biometric and Demographic Pattern Mining in Electrocardiographic Signals: A Study with Machine Learning Algorithms

1st Cristian Veggian Matias
*Postgraduate Program in Electrical Engineering and Industrial Informatics (CPGEI), Federal University of Technology – Paraná (UTFPR)*
Curitiba, Brasil
0009-0008-1281-3099

2nd João Pedro Moreto Lourenção
*Postgraduate Program in Electrical Engineering and Industrial Informatics (CPGEI), Federal University of Technology – Paraná (UTFPR)*
Curitiba, Brasil
email

3rd José Jair Alves Mendes Junior
*Postgraduate Program in Electrical Engineering and Industrial Informatics (CPGEI), Federal University of Technology – Paraná (UTFPR)*
Curitiba, Brasil
email

4th Heitor Silvério Lopes
*Postgraduate Program in Electrical Engineering and Industrial Informatics (CPGEI), Federal University of Technology – Paraná (UTFPR)*
Curitiba, Brasil
email

5th Thiago Henrique Silva
*Postgraduate Program in Electrical Engineering and Industrial Informatics (CPGEI), Federal University of Technology – Paraná (UTFPR)*
Curitiba, Brasil
email

***Abstract*—Electrocardiographic (ECG) signals have emerged as a promising biometric modality due to their physiological uniqueness and resistance to spoofing. This study investigates the feasibility of using ECG-derived features to predict biometric and demographic attributes through supervised and unsupervised machine learning techniques. ECG signals were collected from 26 participants under resting and post-activity conditions using an AD8232-based acquisition system. After preprocessing and heartbeat segmentation, temporal, spectral, statistical, and morphological features were extracted from each cardiac cycle. Supervised models, including k-Nearest Neighbors, Naive Bayes, Random Forest, and Multilayer Perceptron (MLP), were evaluated for classification and regression tasks involving sex, age, weight, height, blood type, and subject identification. Participant-independent validation was adopted for demographic prediction tasks, whereas conventional stratified cross-validation was maintained for closed-set subject identification. In parallel, unsupervised approaches such as k-means, DBSCAN, and hierarchical clustering were applied to investigate the existence of natural groupings in the data. The results showed strong performance for closed-set subject identification, with Random Forest achieving an F1-score of 0.941. However, performance decreased substantially for demographic attributes under participant-independent validation, with the best F1-scores reaching 0.755 for sex and 0.734 for age-group classification. Regression models also showed limited generalization, producing negative coefficients of determination for age, height, and weight prediction. Unsupervised methods did not reveal consistent natural groupings associated with the investigated biometric and demographic characteristics. These findings indicate that ECG-derived features contain strong subject-specific information, while their ability to generalize demographic characteristics to unseen individuals remains limited under the investigated conditions. Larger and more balanced datasets are necessary to improve robustness and evaluate generalization across broader populations.***

***Index Terms*—Biometrics, Biomedical signal processing, Electrocardiography, Feature extraction, Machine learning.**

This work was supported by UTFPR.

## I. Introduction

The use of physiological characteristics for identifying individuals, known as biometrics, has gained increasing relevance in various areas such as security, identity authentication, and medical applications [1], [2]. Among existing biometric approaches, biopotentials—such as electromyographic (EMG), electrocardiographic (ECG), and electroencephalographic (EEG) signals—stand out for reflecting intrinsic characteristics of the human body, potentially unique to each individual [1], [2].

An ECG is the recording of the electrical activity generated by the depolarization of the cardiac muscle during its rhythmic and organized contractions [3]. This electrical activity results directly from the excitation of cardiac cells, whose pattern is influenced by the anatomy of the heart

and individual physiological conditions. Changes in these characteristics can be captured as morphological variations in the ECG waveform [4]–[6]. Thus, differences in cardiac structure and body composition among individuals produce subtly distinct electrocardiographic patterns, which underpins the use of ECG not only for medical diagnosis but also as a possible unique biometric signature for each person [4]–[6].

In parallel, advances in machine learning techniques have provided new tools for analysing and extracting patterns from large volumes of data [4]–[6]. Applications that combine physiological signals with intelligent algorithms have been developed with the aim of exploring complex biometric characteristics that are still not fully understood or leveraged in practical solutions [4]–[6].

In this context, investigating whether demographic attributes can be inferred from ECG-derived features is relevant not only for biometric identification but also for understanding potential sources of demographic bias. If ECG biometric models implicitly exploit information related to sex, age, or body characteristics, their performance may depend on the demographic composition of the training data. Therefore, identifying the demographic information encoded in ECG signals represents an initial step toward the development of more representative datasets and subgroup-aware evaluation protocols. This study does not aim to solve demographic generalization directly, but rather to investigate one of the factors that may influence it.

This work has as its main objective to investigate the potential of surface electrocardiography (ECG) signals in the identification of biometric and demographic patterns using supervised and unsupervised machine learning techniques.

To achieve this goal, the following specific objectives were defined:

- Develop a protocol for acquiring ECG signals under different physiological conditions (rest and light activity), ensuring quality and variability in the recordings;
- Perform the extraction of relevant features from ECG signals, covering temporal, spectral, and morphological information, in order to compose a robust set of attributes for analysis;
- Analyse the ability of unsupervised learning techniques (k-means, DBSCAN, and hierarchical clustering) to identify natural groupings related to biometric and demographic attributes;
- Apply supervised learning algorithms (kNN, Naive Bayes, Random Forest, and MLP) to predict biometric and demographic attributes from ECG signals; and
- Compare the performance between supervised and unsupervised techniques in pattern identification, using appropriate metrics such as F1-Score, $R^2$, AUC, and silhouette coefficient.

## II. Background

Scientific literature indicates that physiological and biomechanical biosignals are highly suitable for biometric identification and authentication [7] [8]. In general, these signals offer three intrinsic advantages over traditional biometric traits: liveness detection, because they reflect active physiological function; resistance to spoofing, because they are dynamic and largely involuntary; and hidden nature, because they cannot be remotely observed as easily as facial images or fingerprints [9]. These properties make biosignals particularly attractive for secure authentication systems.

Among the main modalities, electrocardiography (ECG) is one of the most established. ECG achieves an average recognition rate of 94.3% across five public datasets and is especially valuable because, in addition to being distinctive across individuals, it provides proof of liveness that traditional biometrics cannot guarantee [10] [11]. Electroencephalography (EEG) also shows strong biometric potential, with reported identification accuracy ranging from 80% to 99%, while offering high resistance to replication [12]. Surface electromyography (sEMG) presents competitive performance, reaching an Equal Error Rate (EER) of 3.5% in verification with only 0.8 seconds of gesture data and 90.3% identification accuracy for static wrist and hand gestures. Unlike ECG and EEG, sEMG also enables customizable gestures that can function as "biometric passwords", adding a knowledge-based layer to the authentication process [13]. Photoplethysmography (PPG) likewise shows promising results, with 96% cross-validation accuracy under physical activity, and can reach 99.8% accuracy with an error rate of 0.16% when fused with ECG in a multimodal framework [10].

Multimodal biometric systems further strengthen the case for biosignal-based authentication. The fusion of multiple biosignals, such as ECG, PPG, and respiratory signals, has achieved experimentally validated high accuracy across diverse populations. In addition to improving recognition performance, multimodal systems are harder to spoof and more robust than unimodal approaches, even when sample sizes are limited [9]. However, the benefits of fusion are not always consistent: in some cases, combining ECG and PPG yields only marginal gains over ECG alone, which suggests that the effectiveness of feature-level and decision-level fusion strategies remains insufficiently understood [10] [14].

Despite these advances, several technical barriers still limit practical deployment. Biosignals typically have low amplitude and are highly susceptible to environmental noise and motion artifacts. Conventional electrodes, such as Ag/AgCl, degrade over time and often require skin preparation and conductive gel, which reduces usability in real-world settings. In non-shielded environments, motion artifacts may be up to ten times stronger than the target signal. In addition, signal contamination across modalities remains a major challenge: ECG can interfere with EMG because of overlapping spectral components, and EMG artifacts are particularly difficult to remove from EEG using conventional filtering methods. As a result, signal quality may vary substantially across sessions, locations, and devices, compromising biometric reliability [15] [16] [17].

Methodological fragmentation is another important limitation. Although ECG biometrics have been widely studied, there is still no consensus on standard methodolo-

gies [7] [18] [18]. More broadly, biosignal research still lacks standardized practices for data curation, pretraining objectives, model architectures, and evaluation protocols, unlike fields such as natural language processing and computer vision [19].

These methodological issues are compounded by demographic imbalance and limited dataset documentation. In the PhysioNet repository, fewer than 7% of datasets report the four key demographic variables of race, ethnicity, sex/gender, and age. Among datasets that report sex, most are predominantly male, and most studies have been conducted in North America, especially in the United States. This raises serious concerns about bias in machine learning models trained on such data. Evidence from biometric systems such as facial recognition shows higher false rejection rates for women, children, and racial or ethnic minority groups, as well as higher false-positive rates for these populations. Therefore, the high performance often reported for biosignal biometrics, typically between 80% and 99%, may not generalize consistently across demographic groups [20] [21].

Security and temporal stability also remain open concerns. PPG acquired through non-invasive sensors is vulnerable to spoofing, and many legacy systems still lack integrated safeguards against signals from unverified sources. Signal quality is also affected by factors such as sensor characteristics, emotional state, energy consumption, and the interaction between light and melanin, which may vary across ethnic groups. Although countermeasures such as waveform analysis and liveness detection based on the Hilbert transform have been proposed, standardized defences are still lacking [15]. In addition, some biosignals and biomechanical patterns, such as gait, vary over time, emotional state, and pathological condition. Even within the same individual, physiological biosignals may change from day to day, which complicates robust biometric modelling [15] [8].

Finally, several research directions remain insufficiently explored. Modelling temporal dependencies in PPG is still an unresolved challenge, and although more complex architectures such as bidirectional LSTM networks can better capture temporal structure, they are also more prone to overfitting, sometimes reaching 100% training accuracy with only marginal gains in practical generalization [15]. The integration of physiological biosignals with behavioural signals, such as gait and gestures, is also still at an early stage [14]. At the data level, collection protocols vary widely in terms of activity, duration, environment, and participant instructions, and there are still no clear guidelines for gathering data across broad age ranges, such as 18 to 65 years, or across different physiological conditions [19]. This lack of standardization is aggravated both by data scarcity in some scenarios, which favours overfitting in deep learning models, and by the absence of consistent standards for data organization and storage, even when large datasets are available. As a result, inconsistencies may arise even within the same laboratory, generating rework and reducing reproducibility [21].

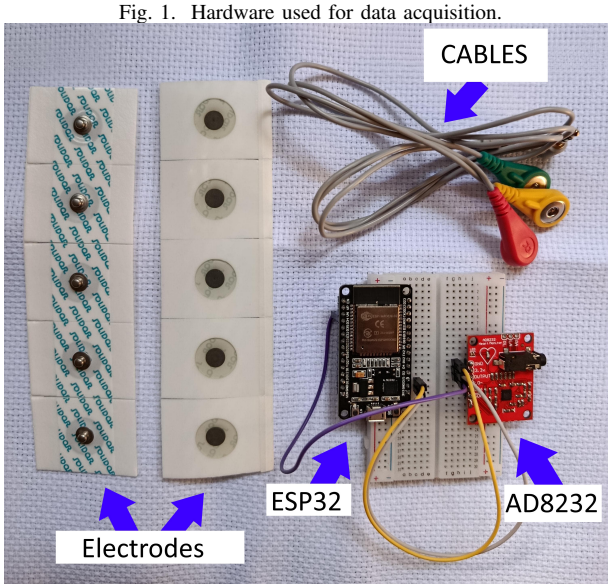
## III. Methodology

This section describes the complete experimental workflow adopted in this study, including ECG acquisition, signal preprocessing, feature extraction, and the machine learning methodologies employed for both supervised and unsupervised analyses. The proposed pipeline was designed to evaluate whether ECG-derived features contain sufficient information to identify biometric and demographic characteristics while also investigating the existence of natural latent structures within the data.

### A. Data Acquisition Hardware

The system used for surface electrocardiography signal acquisition consists of an AD8232-based biopotential acquisition module, an ESP32 microcontroller, and disposable silver/silver chloride (Ag/AgCl) electrodes with conductive gel, as shown in Figure 1. The AD8232 is an integrated biopotential amplifier specifically designed for high-quality, low-noise electrocardiographic signal acquisition, and it is widely employed in portable and wearable applications [22].

Fig. 1. Hardware used for data acquisition.




The signal amplified by the AD8232 was transmitted to the ESP32 microcontroller, which was responsible for analog-to-digital conversion and for sending the data via a serial interface for storage and subsequent analysis. The ESP32 was chosen due to its versatility, low power consumption, and wireless communication capabilities, although in this experiment it was primarily used in wired mode.

The electrodes used were of the Ag/AgCl type, which are standard in electrophysiological signal acquisition due to their low contact impedance and good chemical stability. The use of conductive gel aimed to reduce the resistance between the skin and the electrode, thereby improving the quality of the recorded signal.

### B. Data Collection Protocol

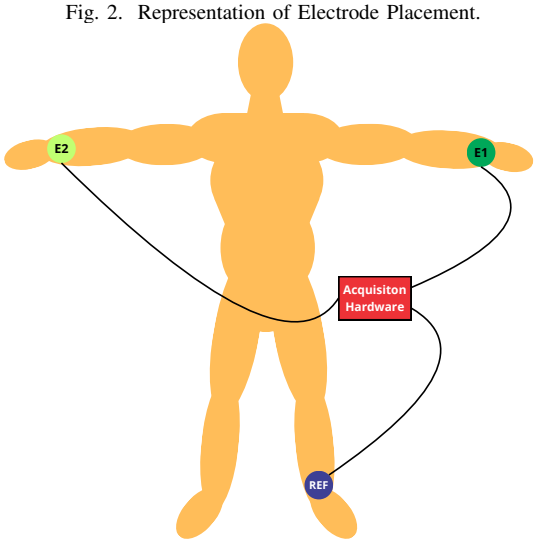
The data collection protocol was designed with the objective of evaluating the possibility of identifying individuals or groups of individuals from surface electrocardiography (ECG) signals. Participant selection followed predefined eligibility criteria intended to reduce major physiological and acquisition-related confounding factors, ensure adequate electrode–skin contact, and protect participants during the post-activity protocol. These criteria were adopted because the present investigation was designed as an initial feasibility study under controlled conditions rather than as a validation for commercial deployment. Individuals aged between 18 and 65 years, of both sexes, and capable of understanding and following the provided instructions were included. These criteria aimed to avoid anatomical interferences in surface signals and to ensure participant cooperation during the procedures.

Individuals with reduced motor mobility, those using medications that affect the nervous or cardiovascular systems, those with wounds, allergies, or dermatitis at the electrode placement sites, those who had consumed alcohol within the previous 24 hours, or those who were pregnant were excluded from the study. These conditions could compromise signal quality or pose ethical and physiological risks. Although these criteria improved experimental control, they reduced the clinical and physiological diversity of the sample and consequently limited the external validity of the results.

Before the start of data collection, all participants were properly instructed regarding the procedures to be performed. The electrode contact areas were cleaned with 70% ethyl alcohol, and participants were instructed to remove metallic objects such as earrings, bracelets, and rings, as well as to avoid carrying electronic devices during recording, such as mobile phones or smartwatches.

The electrodes were positioned on both wrists of the participants, while a reference electrode was placed on the left ankle, as shown in Figure 2.

Fig. 2. Representation of Electrode Placement.




The data collection procedure consisted of two distinct stages for each participant. In the first stage, the signal was recorded under resting conditions, with the individual in a basal state and heart rate within expected average values. In the second stage, data collection was performed immediately after light physical activity, such as jumping jacks or rapidly ascending and descending stairs, in order to increase heart rate. This strategy aimed to obtain signals under different physiological conditions, enabling comparative analyses between resting and active states. The acquired data were properly stored in digital files for subsequent processing and analysis.

### C. Feature Extraction

ECG preprocessing and feature extraction were performed using a reproducible computational pipeline. First, baseline wander was reduced using median filtering. A fourth-order Butterworth band-pass filter with cutoff frequencies of 0.5 and 50 Hz was then applied to attenuate low- and high-frequency interference while preserving the frequency range relevant to ECG analysis. The complete source code, processing parameters, software dependencies, and execution instructions are available at https://github.com/CristianVeggian/ecg-analise-biometrica. Initially, the raw signals underwent a preprocessing stage, which consisted of applying median filters to remove baseline wander and low-frequency components associated with the QRS complex and the T wave.

Subsequently, the filtered signals were segmented into individual heartbeats. For this purpose, the `ecg.ecg()` function from the *biosppy* library was used for automatic R-peak detection. The signal was segmented at each heartbeat by extracting a temporal window of 0.2 seconds before and 0.4 seconds after the corresponding R-peak, allowing the isolation of complete cardiac events for analysis.

Feature extraction was carried out in two stages. First, time-domain, spectral, and statistical features were automatically extracted from each window using the *TSFEL* (*Time Series Feature Extraction Library*), covering a diverse set of numerical descriptors. In addition, specific features related to the morphology of the QRS complex were manually computed, such as the amplitude of the R peak, the amplitude of the S valley, the ratio between these amplitudes ($R/|S|$), the duration of the QRS complex (in seconds), the slope between the R and S points, and the area under the QRS complex, obtained through numerical integration. Table I summarizes the main feature categories considered in this study. Due to the large number of extracted features, the complete feature list, including the corresponding computational definitions, is provided in the project repository.

A total of 162 ECG-derived features were considered, including six QRS-based morphological features and 156 statistical, temporal, spectral, time–frequency, and cepstral descriptors. All extracted features were organized into a CSV file, which also included the corresponding subject identifier and the condition under which the signal was recorded, categorized into two distinct states: rest or activity.

TABLE I
SUMMARY OF ECG-DERIVED FEATURE CATEGORIES USED IN THIS STUDY.

| Category | Examples of extracted features | Description |
|---|---|---|
| Morphological | R amplitude, S amplitude, R/\|S\| ratio, QRS duration, R–S slope, QRS area | Features specifically derived from the morphology of the QRS complex, describing amplitudes, duration, shape, and area of individual heartbeats. |
| Statistical | Mean, median, variance, standard deviation, root mean square, skewness, kurtosis, interquartile range, mean absolute deviation, median absolute deviation | Statistical descriptors characterizing the distribution, dispersion, central tendency, and shape of ECG amplitude values within each heartbeat segment. |
| Temporal | Absolute energy, area under the curve, autocorrelation, zero-crossing rate, peak-to-peak distance, positive and negative turning points, neighbourhood peaks, signal distance | Time-domain descriptors characterizing signal variation, energy, temporal dynamics, transitions, and waveform behavior within each segmented heartbeat. |
| Spectral | Fundamental frequency, median frequency, maximum frequency, spectral centroid, spectral entropy, spectral spread, spectral slope, spectral skewness, spectral kurtosis, spectral roll-on, spectral roll-off, power bandwidth | Frequency-domain descriptors characterizing the distribution, shape, complexity, and concentration of the ECG signal energy across frequencies. |
| Time–frequency | Wavelet energy, wavelet entropy, wavelet absolute mean, wavelet standard deviation, wavelet variance, spectrogram mean coefficients | Time–frequency descriptors representing variations in ECG frequency content over time at different frequency scales. |
| Cepstral | MFCC coefficients, LPCC coefficients | Cepstral-domain descriptors representing compact transformations of the signal spectrum and its spectral envelope. |

### *D. Unsupervised Learning*

Clustering analyses were performed using the Orange Data Mining platform. The process began with the import of files containing the features extracted from each heartbeat along with the metadata file, which included demographic characteristics of the individuals (ID, sex, state, weight, height, blood type). After merging the files and performing initial preprocessing—such as discretizing continuous variables into categories and selecting relevant columns — an outlier filtering step was applied to reduce potential interference in the clustering process.

Three main clustering approaches were used: *k-means*, DBSCAN, and hierarchical clustering. The *k-means* algorithm was configured with the number of clusters adjusted according to the objective of the analysis. In one approach, the goal was to form a number of clusters corresponding to the number of individuals in the dataset, in order to evaluate whether the extracted features would be sufficient to differentiate heartbeats from different individuals, functioning as a biometric signature. In other approaches, the clusters were adjusted to reflect specific characteristics of individuals, such as age group (young, adult, elderly), sex (male or female), weight (three categories), and height (three categories), as well as attempts to cluster by blood type.

In addition to *k-means*, the DBSCAN algorithm was employed to identify clusters with non-spherical shapes and to automatically detect outliers. Hierarchical clustering was applied using the distance matrix between samples, allowing visual analysis of group formation through dendrograms. To evaluate clustering quality, visualizations such as scatter plots and t-SNE projections were used, along with the silhouette metric to assess cluster cohesion and separation.

This approach enabled the evaluation of the ability of features extracted from heartbeats to discriminate individuals or subgroups based on biometric and demographic attributes.

### *E. Supervised Learning*

The supervised learning stage was also conducted using the Orange platform, based on datasets containing features extracted from heartbeats. After importing and organizing the CSV files, preprocessing steps were performed, including file merging, variable discretization, and selection of relevant columns.

The main objective was to evaluate the predictive capability of the extracted features with respect to biometric and demographic attributes of individuals, such as age, sex, weight, height, and blood type. For categorical attributes, including subject ID, sex, recording state, and blood type, supervised classification algorithms were applied. In addition, age, weight, and height were discretized into predefined intervals and evaluated as classification tasks. In this context, four main classifiers were tested: k-Nearest Neighbors (kNN), Naive Bayes, Random Forest, and Multilayer Perceptron. Each algorithm was executed with its default settings, allowing an initial comparison between probabilistic approaches, instance-based methods, and ensemble learning techniques.

For numerical variables (age, weight, and height), two distinct strategies were evaluated. In the first approach, these variables were discretized into predefined intervals to transform the problem into a classification task. However, due to class imbalance observed in some variables—especially age—experiments were also conducted using supervised regression to predict continuous values. For regression tasks, the algorithms used included k-Nearest Neighbors, Linear Regression, Random Forest, and Multilayer Perceptron.

Model validation was performed using the Test and Score widget. For demographic classification and regression tasks, participant-independent validation was adopted using cross-validation grouped by subject identifier (ID). Therefore, all heartbeat segments belonging to the same participant were assigned to the same fold, preventing samples from a given individual from appearing simultaneously in the training and test

sets. For closed-set subject identification, conventional stratified 10-fold cross-validation was maintained, since subject-wise splitting would remove from the training set the identity class to be predicted. Classification performance was evaluated using accuracy, precision, recall, F1-score, AUC, and Matthews correlation coefficient (MCC), whereas regression performance was assessed using MSE, RMSE, MAE, MAPE, and $R^2$.

## IV. Results

This section presents the experimental results obtained from both the unsupervised and supervised learning analyses. Initially, the characteristics of the generated dataset are described, followed by the evaluation of clustering algorithms and predictive models. The results are analyzed using quantitative performance metrics and visual inspection in order to assess the feasibility of identifying biometric and demographic information from ECG-derived features.

### A. Generated Dataset

A total of 52 recordings were performed across 26 distinct participants. In total, 1,858 electrocardiographic segments were obtained, distributed between resting and activity conditions.

The generated dataset can be found in the following GitHub repository: https://github.com/CristianVeggian/ecg-analise-biometrica.

Figure 3 shows the distribution of the dataset metadata, which were used as target variables in classification and regression tasks. A significant imbalance can be observed in some variables, such as sex and age. The number of segments from male participants is substantially higher compared to female participants, which may impact the performance of supervised classifiers. Similarly, there is a predominance of participants classified in the "young" age group, with lower representation in the "adult" and "elderly" categories.

The distributions of weight and height show a more balanced division among the established categories, which is favorable for model training. In the case of blood type, there is a concentration of records in the "Not Specified" (N/S) category, in addition to a non-uniform distribution among the different types, representing an additional challenge for predictive algorithms.

These observations highlight the need to carefully evaluate the impact of imbalance during the supervised learning stage, especially in multiclass classification tasks. For discretized numerical variables, the imbalance motivated additional experiments using supervised regression in order to explore alternative approaches and mitigate the limitations imposed by the data distribution.

### B. Unsupervised Learning

Three main unsupervised learning algorithms were tested: k-means, DBSCAN, and hierarchical clustering. The quantitative evaluation of clustering quality was performed using the silhouette coefficient.

The elbow method was applied to determine the optimal number of clusters in the k-means algorithm, testing values of $k$ between 2 (minimum expected number of clusters) and 26 (maximum expected, based on the number of individuals). Despite this approach, it was not possible to identify a clear inflection point in the curve, reinforcing the absence of well-defined clustering structures in the dataset, as shown in Figure 4.

Furthermore, even with manual adjustments of $k$, the silhouette coefficient values remained low, indicating that the separation between clusters did not represent cohesive or consistent groupings. These results suggest that, within the context of this study and given the extracted features, ECG signals did not exhibit sufficient natural separations for meaningful clustering through unsupervised learning.

Despite attempts to optimize the DBSCAN algorithm through systematic tuning of the *eps* and *min_samples* parameters, the resulting partition was highly unbalanced. DBSCAN identified two clusters, containing 1,650 and 20 samples, respectively, while only two samples (0.1%) were classified as noise. The silhouette coefficient, calculated excluding noise samples, was 0.520. Although this value indicates moderate separation between the identified clusters, the strong imbalance in cluster sizes indicates that the data were predominantly represented by a single large cluster rather than by two comparably sized natural groups.

Additionally, the t-SNE projection did not reveal a clear correspondence between the DBSCAN partition and the investigated metadata. When the same projection was visualized according to recording state, sex, and subject ID, substantial overlap remained between the corresponding groups. Therefore, although DBSCAN produced a moderately separated partition according to the silhouette coefficient, the identified clusters could not be clearly associated with the investigated biometric or demographic characteristics.

Figure 5 presents the t-SNE projection of the DBSCAN results together with the corresponding metadata. Cluster 0 represents the dominant group containing 1,650 samples, Cluster 1 contains 20 samples, and only two samples were classified as noise. The comparison between the clustering result and the available metadata illustrates the absence of a clear demographic or biometric interpretation for the partition obtained by DBSCAN.

Hierarchical clustering was also applied using Euclidean distance and the *Ward* linkage method. The number of clusters was not defined a priori, since this analysis was exploratory and aimed at investigating the intrinsic organization of the ECG-derived feature space. The *Ward* criterion was selected because it minimizes within-cluster variance and is appropriate for continuous numerical data analyzed with Euclidean distances.

Inspection of the dendrogram revealed two major branches separated at a relatively high linkage distance, motivating the evaluation of a two-cluster partition. The resulting partition contained 1,038 samples in Cluster 1 and 634 samples in Cluster 2. Although the dendrogram revealed a clear high-

Fig. 3. Distribution of participants according to (a) state, (b) sex, (c) age group, (d) weight group, (e) height group, and (f) blood type.

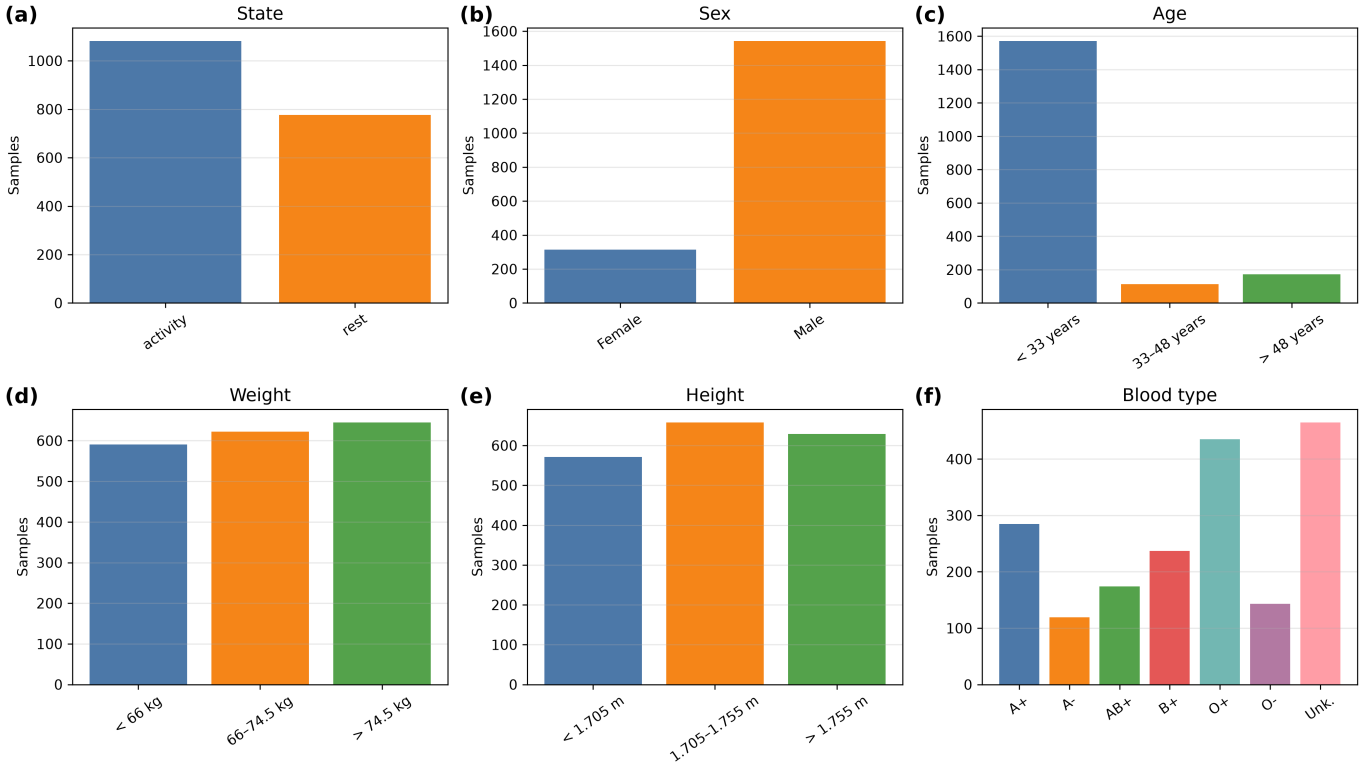



Fig. 4. Elbow method applied to k-means.

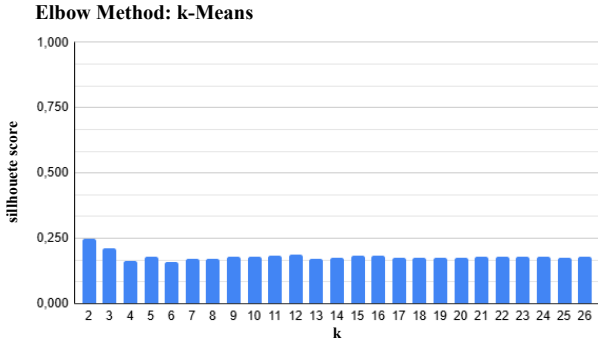
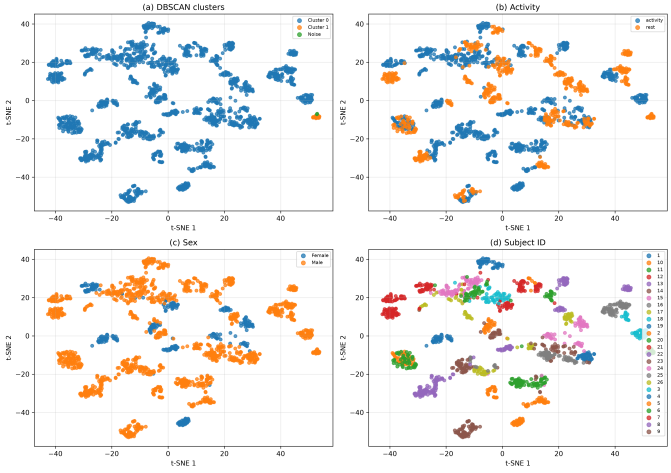


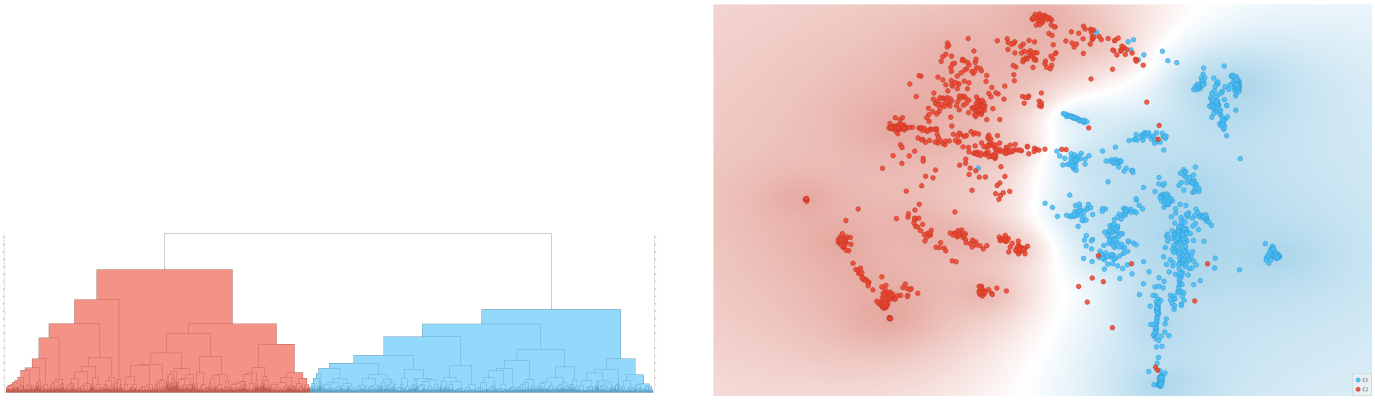

Fig. 5. t-SNE visualization of the DBSCAN clustering results and selected metadata. Cluster 0 and Cluster 1 represent the groups identified by DBSCAN, whereas Noise represents samples that were not assigned to any cluster. The remaining panels show the same t-SNE projection colored according to recording state, sex, and subject ID.




level bifurcation, the silhouette analysis showed substantial variability in cluster cohesion, including negative silhouette values for several samples. This indicates that a considerable number of observations were not clearly separated between the two clusters.

Figure 6 presents the dendrogram together with the corresponding t-SNE projection. The t-SNE visualization showed considerable overlap between the two groups and did not reveal a clear correspondence with the investigated demographic or biometric attributes. Therefore, the two-cluster solution was interpreted as an exploratory representation of the hierarchical structure of the feature space rather than as evidence of two predefined or well-separated demographic or biometric groups.

These findings indicate that, similarly to *k-means* and DBSCAN, hierarchical clustering did not identify natural groups that could be consistently associated with the investigated demographic or biometric attributes. The presence of two major branches in the dendrogram reflects a high-level hierarchical



(a) Hierarchical clustering dendrogram.

(b) t-SNE projection of the two-cluster partition.

Fig. 6. Hierarchical clustering analysis using Ward linkage and Euclidean distance. (a) Dendrogram showing the hierarchical organization of the ECG-derived samples; the two-cluster partition corresponds to the two major branches observed before the final high-level merge. (b) t-SNE projection of the resulting two-cluster partition, showing regions of overlap between the clusters. The partition contained 1,038 samples in Cluster 1 and 634 samples in Cluster 2.

division of the feature space, but the silhouette distribution and the overlapping t-SNE projection indicate that this division should not be interpreted as two clearly separable biological or demographic populations.

### C. Supervised Learning

Table II presents the F1-scores obtained for the supervised classification tasks. A clear distinction can be observed between closed-set subject identification and participant-independent demographic classification.

TABLE II
F1-SCORE BY ATTRIBUTE AND ALGORITHM.

| Class | kNN | Naive Bayes | Random Forest | MLP |
|---|---|---|---|---|
| ID (26 classes) | 0.343 | 0.834 | **0.941** | 0.922 |
| Sex (2 classes) | **0.755** | 0.661 | 0.696 | 0.730 |
| Age (3 classes) | **0.734** | 0.615 | 0.721 | 0.713 |
| State (rest/activity) | 0.482 | **0.617** | 0.551 | 0.568 |
| Height (3 classes) | **0.291** | 0.175 | 0.157 | 0.145 |
| Weight (3 classes) | 0.192 | 0.203 | 0.195 | **0.280** |
| Blood type (7) | 0.167 | 0.204 | 0.173 | **0.229** |

For closed-set subject identification, Random Forest achieved the highest F1-score (0.941), followed by MLP (0.922) and Naive Bayes (0.834), whereas kNN obtained a considerably lower F1-score of 0.343. These results indicate that the extracted ECG features contain strong subject-specific information when the identities evaluated during testing are represented in the training data.

In contrast, performance decreased when participant-independent validation was applied to demographic attributes. Sex classification achieved its highest F1-score with kNN (0.755), followed by MLP (0.730), Random Forest (0.696), and Naive Bayes (0.661). For age-group classification, kNN also obtained the highest F1-score (0.734), while Random Forest and MLP achieved 0.721 and 0.713, respectively. State classification presented lower performance, with Naive Bayes achieving the highest F1-score of 0.617.

The remaining demographic attributes presented considerably lower predictive performance. For height classification, the highest F1-score was 0.291 with kNN. Weight classification reached a maximum F1-score of 0.280 with MLP, whereas blood-type classification reached a maximum F1-score of 0.229, also with MLP. These results indicate that, under participant-independent validation, the extracted ECG features provided limited discriminative information for these attributes.

Overall, the reduction in demographic classification performance compared with closed-set subject identification suggests that ECG-derived features contain substantial subject-specific information, whereas demographic patterns are considerably more difficult to generalize to participants not represented during model training.

Within the supervised learning analysis, regression experiments were also conducted to evaluate the ability of the models to predict continuous demographic variables, namely height, age, and weight, under participant-independent validation.

Table III presents the performance of the regression models for height prediction. All evaluated models produced negative coefficients of determination, indicating limited generalization to unseen participants. Among the evaluated methods, kNN achieved the highest $R^2$ value ($R^2 = -0.416$), with an RMSE of 0.082 and an MAE of 0.063. Random Forest obtained $R^2 = -0.619$, while Linear Regression and MLP presented substantially lower values of $R^2 = -2.268$ and $R^2 = -7.854$, respectively. These results indicate that none of the evaluated models reliably explained height variability for participants not represented during training.

TABLE III
PERFORMANCE OF REGRESSION MODELS FOR HEIGHT PREDICTION.

| Model | MSE | RMSE | MAE | MAPE (%) | $R^2$ |
|---|---|---|---|---|---|
| kNN | 0.007 | 0.082 | 0.063 | 3.582 | **-0.416** |
| Linear Regression | 0.015 | 0.124 | 0.081 | 4.609 | -2.268 |
| Neural Network (MLP) | 0.042 | 0.204 | 0.147 | 8.429 | -7.854 |
| Random Forest | 0.008 | 0.087 | 0.067 | 3.812 | -0.619 |

Table IV presents the regression results for age prediction. Similar to the height prediction task, all models produced

negative $R^2$ values. Random Forest achieved the highest coefficient of determination ($R^2 = -0.497$), with an RMSE of 13.369 and an MAE of 9.064 years. kNN obtained a similar performance, with $R^2 = -0.594$, whereas MLP and Linear Regression presented $R^2$ values of -0.848 and -3.520, respectively. Therefore, none of the evaluated models outperformed a baseline based on the mean target value when predicting the age of unseen participants.

TABLE IV
PERFORMANCE OF REGRESSION MODELS FOR AGE PREDICTION.

| Model | MSE | RMSE | MAE | MAPE (%) | $R^2$ |
|---|---|---|---|---|---|
| kNN | 190.320 | 13.796 | 9.632 | 33.391 | -0.594 |
| Linear Regression | 539.767 | 23.233 | 12.574 | 47.098 | -3.520 |
| Neural Network (MLP) | 220.725 | 14.857 | 10.263 | 36.180 | -0.848 |
| Random Forest | 178.719 | 13.369 | 9.064 | 30.520 | **-0.497** |

Table V presents the results obtained for weight prediction. Random Forest achieved the best performance among the evaluated models, with an RMSE of 15.991 kg, an MAE of 11.680 kg, and $R^2 = -0.369$. However, the negative coefficient of determination indicates that its predictions did not generalize adequately to unseen participants. kNN obtained $R^2 = -0.456$, while MLP and Linear Regression achieved $R^2 = -0.807$ and $R^2 = -1.162$, respectively.

TABLE V
PERFORMANCE OF REGRESSION MODELS FOR WEIGHT PREDICTION.

| Model | MSE | RMSE | MAE | MAPE (%) | $R^2$ |
|---|---|---|---|---|---|
| kNN | 272.097 | 16.495 | 12.339 | 16.500 | -0.456 |
| Linear Regression | 403.913 | 20.098 | 14.904 | 20.138 | -1.162 |
| Neural Network (MLP) | 337.636 | 18.375 | 14.362 | 19.132 | -0.807 |
| Random Forest | 255.714 | 15.991 | 11.680 | 15.337 | **-0.369** |

Taken together, the regression results demonstrate that the evaluated models were unable to reliably generalize continuous demographic attributes across participants. This contrasts with the strong closed-set subject identification performance and further indicates that subject-specific ECG characteristics are more readily captured by the extracted features than demographic patterns that generalize across individuals.

## V. DISCUSSION

The results obtained in this study reveal an important distinction between closed-set biometric identification and demographic prediction from ECG-derived features. While subject identification remained highly effective, participant-independent validation substantially reduced the performance of models applied to demographic attributes. This finding indicates that ECG-derived features contain strong subject-specific information, but their ability to generalize demographic characteristics to previously unseen individuals is considerably more limited.

For closed-set subject identification, Random Forest achieved the highest performance, with an F1-score of 0.941, followed by the Multilayer Perceptron with an F1-score of 0.922. In this task, conventional stratified cross-validation was maintained because all identity classes must be represented in the training set. These results reinforce the hypothesis that morphological, temporal, statistical, and spectral characteristics extracted from ECG signals contain discriminative information capable of differentiating individuals.

In contrast, when participant-independent validation was adopted for demographic classification, performance decreased substantially. The best F1-scores were 0.755 for sex classification and 0.734 for age-group classification, both obtained with kNN. State classification reached a maximum F1-score of 0.617 with Naive Bayes, whereas height, weight, and blood-type classification presented considerably lower results. These findings suggest that part of the high performance observed under heartbeat-level cross-validation was likely associated with subject-specific patterns shared between training and test samples rather than with demographic characteristics that generalize across individuals.

The regression experiments provide further evidence of this limitation. Under participant-independent validation, all evaluated models produced negative $R^2$ values for age, height, and weight prediction. The best results were $R^2$ = -0.416 for height using kNN, $R^2$ = -0.497 for age using Random Forest, and $R^2$ = -0.369 for weight using Random Forest. Negative coefficients of determination indicate that, under the investigated conditions, the models did not generalize these continuous demographic attributes to unseen participants and performed worse than a baseline prediction based on the target mean.

These results highlight the methodological importance of subject-wise validation when multiple heartbeat segments are extracted from the same participant. Because several segments originate from a single individual, conventional sample-level cross-validation may allow subject-specific physiological information to appear in both training and test sets, resulting in optimistic estimates of model performance. The participant-independent evaluation therefore provides a more conservative and realistic assessment of the demographic information that can be generalized from ECG-derived features.

The unsupervised learning results are consistent with this interpretation. k-means, DBSCAN, and hierarchical clustering did not reveal natural groupings that could be clearly associated with the investigated demographic and biometric characteristics. In particular, DBSCAN identified two highly unbalanced clusters, containing 1,650 and 20 samples, respectively, with only two samples (0.1%) classified as noise. Although the silhouette coefficient of 0.520 indicated moderate separation between the DBSCAN clusters, their substantial size imbalance and the absence of a clear correspondence with the available metadata suggest that this partition does not represent a meaningful demographic or biometric grouping. Similarly, the other clustering approaches showed considerable overlap between groups and no consistent association with the investigated attributes. Nevertheless, the strong closed-set identification results demonstrate that the absence of clear unsupervised groupings does not imply an absence of subject-specific information in the ECG signals.

Several factors may have contributed to the limited de-

mographic generalization observed in this study. The dataset contains only 26 participants and presents demographic imbalance, particularly regarding sex, age, and blood type. In addition, the relatively small number of subjects restricts the diversity represented in participant-independent validation folds. High feature dimensionality and physiological variability between recording conditions may also contribute to the difficulty of learning patterns that generalize across participants.

Therefore, the results should be interpreted as evidence that ECG-derived features are highly informative for distinguishing individuals already represented in the biometric system, while their ability to predict demographic attributes for unseen individuals remains limited under the investigated conditions. Larger and more representative datasets, together with participant-independent evaluation protocols, will be necessary to determine whether demographic information can be reliably generalized from ECG signals.

## VI. Conclusion

This work investigated the feasibility of extracting biometric and demographic information from ECG signals using supervised and unsupervised machine learning techniques. A dataset containing ECG recordings from 26 participants under resting and post-activity conditions was collected, and more than 160 temporal, spectral, statistical, and morphological features were extracted from each heartbeat.

The participant-independent evaluation revealed a clear distinction between biometric identification and demographic prediction. ECG-derived features demonstrated strong discriminative capability for closed-set subject identification, with Random Forest achieving an F1-score of 0.941. However, when heartbeat segments from the same participant were prevented from appearing simultaneously in the training and test sets, classification performance decreased substantially for demographic attributes.

Regression results further demonstrated limited generalization to unseen participants. All evaluated models produced negative $R^2$ values for age, height, and weight prediction, indicating that the extracted features were insufficient to reliably estimate these continuous demographic variables under participant-independent validation.

Similarly, unsupervised approaches, including k-means, DBSCAN, and hierarchical clustering, were unable to identify consistent latent structures associated with the investigated biometric and demographic attributes. Overall, these findings indicate that ECG-derived features contain strong subject-specific information, supporting their potential for closed-set biometric identification, while demographic characteristics were not reliably generalized from the extracted feature space in the present dataset.

The results also emphasize the importance of participant-independent validation in studies involving multiple physiological segments from the same individuals, since sample-level cross-validation may lead to optimistic performance estimates due to subject-specific information shared between training and test sets.

Future work will focus on expanding the dataset with a larger and more demographically diverse cohort, improving the balance among demographic groups, investigating feature selection and dimensionality reduction techniques, and evaluating additional machine learning and deep learning approaches under participant-independent validation protocols.

## References


[1] D. Wang, Y. Si, W. Yang, G. Zhang, and J. Li, "A novel electrocardiogram biometric identification method based on temporal-frequency autoencoding," *Electronics*, vol. 8, no. 6, 2019.

[2] A. N. Uwaechia and D. A. Ramli, "A comprehensive survey on ecg signals as new biometric modality for human authentication: Recent advances and future challenges," *IEEE Access*, vol. 9, pp. 97760–97802, 2021.

[3] K. Najarian and R. Splinter, *Biomedical Signal and Image Processing*. Boca Raton, FL: CRC Press, 2 ed., May 2012.

[4] T. S. Lugovaya, "Biometric human identification based on electrocardiogram," master's thesis, Faculty of Computing Technologies and Informatics, Electrotechnical University "LETI", Saint-Petersburg, Russian Federation, June 2005.

[5] T. W. Shen, W. J. Tompkins, and Y. H. Hu, "One-lead ECG for identity verification," in *Proceedings of the Second Joint 24th Annual Conference and the Annual Fall Meeting of the Biomedical Engineering Society] [Engineering in Medicine and Biology*, IEEE, 2003.

[6] L. Biel, O. Pettersson, L. Philipson, and P. Wide, "ECG analysis: a new approach in human identification," *IEEE Trans. Instrum. Meas.*, vol. 50, pp. 808–812, June 2001.

[7] T. M. C. Pereira, R. C. Conceição, V. Sencadas, and R. Sebastião, "Biometric recognition: A systematic review on electrocardiogram data acquisition methods," *Sensors (Basel)*, vol. 23, p. 1507, Jan. 2023.

[8] P. Connor and A. Ross, "Biometric recognition by gait: A survey of modalities and features," *Comput. Vis. Image Underst.*, vol. 167, pp. 1–27, Feb. 2018.

[9] A. Hadid, N. Evans, S. Marcel, and J. Fierrez, "Biometrics systems under spoofing attack: An evaluation methodology and lessons learned," *IEEE Signal Process. Mag.*, vol. 32, pp. 20–30, Sept. 2015.

[10] Y. Che, L. Du, G. Tang, and S. Ling, "A biometric identification for multi-modal biomedical signals in geriatric care," *Sensors*, vol. 24, p. 6558, Oct. 2024.

[11] N. Belgacem, R. Fournier, A. Nait-Ali, and F. Bereksi-Reguig, "A novel biometric authentication approach using ecg and emg signals," *Journal of Medical Engineering amp; Technology*, vol. 39, p. 226–238, Apr. 2015.

[12] S. Zhang, L. Sun, X. Mao, C. Hu, and P. Liu, "Review on EEG-based authentication technology," *Comput. Intell. Neurosci.*, vol. 2021, p. 5229576, Dec. 2021.

[13] J. He and N. Jiang, "Biometric from surface electromyogram (sEMG): Feasibility of user verification and identification based on gesture recognition," *Front. Bioeng. Biotechnol.*, vol. 8, p. 58, Feb. 2020.

[14] A. R. Mishra, R. Kumar, V. Gupta, S. Prabhu, R. Upadhyay, P. C. Chhipa, S. Rakesh, H. Mokayed, D. Das Chakladar, K. De, M. Liwicki, F. Simistira Liwicki, and R. Saini, "SignEEG v1.0: Multimodal dataset with electroencephalography and hand-written signature for biometric systems," *Sci. Data*, vol. 11, p. 718, July 2024.

[15] D. Aguilar, A. Martínez-Cruz, K. A. Ramírez-Gutiérrez, and M. Morales-Sandoval, "Ppg-based biometric authentication: A review on architectures, datasets, attacks and security challenges," *Computational and Structural Biotechnology Journal*, vol. 28, p. 511–528, Jan. 2025.

[16] K.-M. Chang, P.-T. Liu, and T.-S. Wei, "Electromyography parameter variations with electrocardiography noise," *Sensors (Basel)*, vol. 22, p. 5948, Aug. 2022.

[17] H. Yu, M. Ma, B. Zhang, A. Wang, G. Zhong, Z. Zhou, C. Liu, C. Li, J. Fang, Y. He, D. Ren, F. Deng, Q. Hong, Y. Zhao, and X. Guo, "Bionic sensors for biometric acquisition and monitoring: Challenges and opportunities," *Sensors (Basel)*, vol. 25, p. 3981, June 2025.

[18] L. Xu, M. Xu, Y. Ke, X. An, S. Liu, and D. Ming, "Cross-dataset variability problem in EEG decoding with deep learning," *Front. Hum. Neurosci.*, vol. 14, p. 103, Apr. 2020.

[19] X. Gu, Y. Shu, J. Han, Y. Liu, Z. Liu, J. Anibal, V. Sangha, E. Phillips, B. Segal, H. Yuan, F. Liu, K. Branson, P. Schwab, D. Belgrave, L. Clifton, D. Spathis, V. Lampos, A. A. Faisal, and D. A. Clifton, "Foundation models for biosignals: A survey," *Techrxiv*, Aug. 2025.

[20] K. Kotwal and S. Marcel, "Review of demographic fairness in face recognition," *IEEE Trans. Biom. Behav. Identity Sci.*, vol. 8, pp. 20–45, Jan. 2026.

[21] S. Jiang, P. Ashar, M. M. H. Shandhi, and J. Dunn, "Demographic reporting in biosignal datasets: a comprehensive analysis of the PhysioNet open access database," *Lancet Digit. Health*, vol. 6, pp. e871–e878, Nov. 2024.

[22] J. J. A. Mendes Junior, D. P. Campos, L. C. d. A. V. D. Biassio, P. C. Passos, P. B. Júnior, A. E. Lazzaretti, and E. Krueger, "Ad8232 to biopotentials sensors: Open source project and benchmark," *Electronics*, vol. 12, no. 4, 2023.