Automated Detection of Chagas Disease via CatBoost

Rogério Ferreira Rodrigues Júnior
Departamento de Estatística e Matemática Aplicada
Universidade Federal do Ceará
Fortaleza, Brasil
rger0904@hotmail.com

João Gabriel Soares Ferreira
Departamento de Computação
Universidade Federal do Ceará
Fortaleza, Brasil
gabrielsoares@alu.ufc.br

João Samuel Maciel de Sales
Departamento de Computação
Universidade Federal do Ceará
Fortaleza, Brasil
joaosamuel.2001@gmail.com

Marcio Barros Oliveira de Souza
Departamento de Estatística e Matemática Aplicada
Universidade Federal do Ceará
Fortaleza, Brasil
marciobarros.2310@gmail.com

Gisela V. Clemente
Departamento de Electrónica
Universidad Tecnológica Nacional
Buenos Aires, Argentina
gclemente@mate.unlp.edu.ar

Pedro Joás Freitas Lima
Departamento de Estatística e Matemática Aplicada
Universidade Federal do Ceará
Fortaleza, Brasil
pjoas26@gmail.com

João Paulo do Vale Madeiro
Departamento de Computação
Universidade Federal do Ceará
Fortaleza, Brasil
jpaulo.vale@dc.ufc.br

***Abstract*—Chagas disease, caused by *Trypanosoma cruzi*, is a neglected tropical disease that continues to pose a significant health threat in Latin America and is increasingly found in non-endemic regions due to migration. The chronic cardiac form of the disease (chronic Chagasic cardiomyopathy) often develops silently, making early detection difficult and effective treatment challenging. In this work, a machine learning approach based on the CatBoost algorithm is proposed for the detection of Chagas disease using electrocardiogram signals. A large-scale dataset from multiple sources was collected and processed, enabling the extraction of 425 features encompassing demographic and advanced ECG waveform characteristics. Severe class imbalance was addressed through targeted evaluation metrics and segmentation by age. The CatBoost model demonstrated robust performance, especially for individuals aged equal to or younger than 45 years, achieving a ROC-AUC of 0.89 and balanced accuracy of 0.81. Performance decreased to ROC-AUC of 0.81 and balanced accuracy of 0.73 for older individuals ($>$ 45 years). Despite these favorable metrics, the model presented notable limitations, particularly a low Matthews Correlation Coefficient (approximately 0.18) and low precision in predicting the minority class due to severe class imbalance, leading to a substantial number of false positives. This highlights the need for further methodological refinements to improve minority class detection. These findings reinforce the clinical potential of AI-empowered ECG analysis for scalable Chagas screening, particularly in resource-constrained settings.**

***Index Terms*—Chagas disease, ECG, CatBoost, classification.**

I. Introduction

Chagas disease (ChD), caused by the protozoan *Trypanosoma cruzi*, remains a major public health challenge in Latin America, affecting millions of people and causing significant morbidity and mortality due to its chronic cardiac complications. With globalization and migration, new cases have emerged in non-endemic regions, increasing the importance of scalable diagnostic solutions [1], [2]. The disease progresses in two stages: an acute phase, typically asymptomatic or mild, and a chronic phase that may remain silent for years before manifesting as severe cardiac disorders, notably chronic Chagasic cardiomyopathy (CCC) [1]–[3]. The CCC leads to progressive cardiac damage, including conduction abnormalities, arrhythmias, heart failure, and increased risk of sudden cardiac death (SCD), often with devastating consequences [2], [4].

Electrocardiogram (ECG) analysis is a non-invasive and widely available diagnostic tool capable of capturing structural and functional cardiac changes associated with chronic ChD [5]. However, clinical diagnosis remains difficult, particularly in the early stages, due to subtle or nonspecific ECG findings and a prolonged asymptomatic period [6]. In this context, the use of artificial intelligence (AI) and machine learning (ML) models to analyze ECG signals has emerged as a promising avenue to enhance screening, risk stratification, and early diagnosis of ChD [4], [5], [7].

The objective of this work is to develop and validate an automated approach using the CatBoost algorithm, a gradient boosting method well-suited for tabular data and categorical variables, to distinguish individuals with and without ChD based on ECG features [8]. By leveraging a large and heterogeneous dataset, extracting detailed demographic and waveform-based features, and addressing severe class imbalance, the

study seeks to improve both the accuracy and reliability of ChD detection. Data segmentation by age groups, motivated by evidence that age is a key discriminative factor, enables an exploration of model performance across different risk strata. This work contributes to the expanding body of research on AI-powered biomedical diagnostics, aiming to support clinical decision-making and large-scale screening efforts, particularly in regions with limited resources.

II. Literature Review: Categorical boosting in Biomedical Research

Categorical boosting (CatBoost) is a gradient boosting algorithm that natively handles categorical variables and uses ordered boosting to mitigate overfitting [8]. The literature notes that CatBoost often excels on heterogeneous tabular data. With relatively few hyperparameters to tune and efficient training, CatBoost scales well to large datasets on 200k Intensive Care Unit (ICU) records [9]. Crucially, CatBoost models are interpretable: Feature importances and SHapley Additive exPlanations (SHAP) values can highlight biomarkers or clinical predictors. These qualities: accuracy, robustness, and interpretability, have led to CatBoost's widespread adoption in medical ML, and justify its adoption in this study. Some applications of CatBoost in biomedical tasks are discussed below, showing its usefulness to the field.

In cardiology, Reference [10] extracted multiple spectrogram features from phonocardiograms using convolutional neural networks (CNNs) and fed them into CatBoost to noninvasively detect left-ventricular diastolic dysfunction, yielding an AUC of 0.911 and an accuracy of 88.2%. Reference [11] used CatBoost to predict anticancer drug combination synergy on the NCI-ALMANAC dataset and found that CatBoost outperformed deep neural networks, eXtreme Gradient Boosting (XGBoost), and logistic regression across all metrics. CatBoost also achieved the highest accuracy (99%) when predicting survival outcomes in recurrent cervical cancer, surpassing other classifiers [12].

The detection of ChD, particularly in its chronic cardiac form, CCC, has become an increasingly relevant topic in biomedical research due to the importance of early intervention and the often silent progression of the condition. Many patients remain asymptomatic for years, which complicates diagnosis and delays treatment. In this scenario, leveraging artificial intelligence in the analysis of biomedical signals, such as ECG, presents an innovative and practical solution.

In studies specifically on ChD, CatBoost has also started to appear. Reference [13] included CatBoost in a ML ensemble to predict *Trypanosoma cruzi* infection based on demographic and clinical survey data. Reference [14] developed a multiparametric classifier for SCD in Chagas cardiomyopathy patients using ECG and clinical features, testing CatBoost among other algorithms, demonstrating the relevance of CatBoost for Chagas related to cardiac risk prediction. These examples suggest that CatBoost can leverage large patient cohorts and ECG datasets to improve risk stratification in ChD.

III. Theoretical Background

This section briefly explains the theory underlying the methods employed in this work.

A. CatBoost

CatBoost is designed to handle categorical predictors efficiently and remove the prediction shift bias that appears when the same sample is used both to grow the trees and to compute the gradients [8], [15]. Comparative studies report that CatBoost achieves superior area under the curve (AUC) and accuracy on tasks rich in categorical attributes, often outperforming XGBoost and Light Gradient Boosting Machine (LightGBM), although at the cost of longer training times [16]. The main CatBoost innovations are discussed below.

1) Ordered Encoding of Categorical Variables: Before training, the dataset is shuffled several times. For each permutation, every category c of a categorical attribute is replaced with an ordered Counter-Target Rate (CTR):

$$CTR(c) = \frac{curCount(c) + prior}{maxCount + 1}, \tag{1}$$

where $curCount(c)$ is the number of preceding samples whose category equals c, $maxCount$ is the frequency of the most common category, and prior is a hyperparameter because only previous observations are used, the encoding prevents target leakage and reduces variance [8]. Low cardinality values are one-hot encoded, whereas multiple smoothed versions of the CTR are binarized to enrich the feature space.

2) Ordered Boosting and Prediction Shift Correction: CatBoost updates the model by constructing trees based on gradients computed over the entire training dataset. This approach, however, can introduce a systematic bias in the estimates. CatBoost computes the gradient of each instance x_i only with models trained on samples that precede x_i in every permutation, a procedure called ordered boosting. This strategy markedly reduces overfitting, especially on small or noisy datasets [8].

3) Symmetrical Oblivious Trees: Each tree grown by CatBoost is symmetric. All nodes at a given depth split on the same feature using the same threshold. Such oblivious trees are perfectly balanced and have a fixed depth d, which enables very fast inference (as decision tables fit in cache), fewer free parameters (thereby reducing the risk of overfitting), and straightforward graphics processing unit (GPU) parallelization. During the tree construction, many (feature, split) pairs are tested. The candidate tree that maximizes the cosine similarity between the residuals r_i and its predictions $h(x_i)$ is selected. The cosine similarity is defined as:

$$\cos(\theta) = \frac{\sum_i r_i \cdot h(x_i)}{\sqrt{\sum_i r_i^2} \cdot \sqrt{\sum_i h(x_i)^2}}. \tag{2}$$

4) Regularization and Overffitting Detection: Besides the learning rate η and ℓ_2 leaf regularization, CatBoost provides an internal overfitting detector that halts training when the validation loss ceases to improve significantly. Furthermore,

bagging across permutations and the injection of gradient noise further enhance robustness.

B. Class-Imbalanced Learning Metrics

Robust evaluation under severe class imbalance requires metrics that do not inflate performance due to majority class prevalence and that appropriately reward genuine improvements on the minority class [17], [18]. Several widely adopted metrics are particularly suitable when instances of a class are scarce. These metrics are formally defined below.

1) Stratified (Per–Class) Brier Score: Let $y_{ic} \in \{0,1\}$ denote the indicator of instance i belonging to class c and $\hat{p}_{ic}$ its predicted probability [19]. The class-specific Brier score is

$$BS_c = \frac{1}{N_c} \sum_{i:y_{ic}=1} (\hat{p}_{ic} - 1)^2 + \frac{1}{N_c} \sum_{i:y_{ic}=0} \hat{p}_{ic}^2, \tag{3}$$

where N_c is the number of observations of class c. The stratified Brier score (SBS) averages these values over C classes,

$$SBS = \frac{1}{C} \sum_{c=1}^{C} \mathrm{BS}_c, \tag{4}$$

ensuring that every class contributes equally regardless of frequency because the quadratic penalty operates on calibrated probabilities. SBS simultaneously captures discrimination and calibration without being dominated by the majority class.

2) Balanced Accuracy: It is the arithmetic mean of per-class recalls,

$$BA = \frac{1}{C} \sum_{c=1}^{C} \frac{TP_c}{TP_c + FN_c}, \tag{5}$$

where C denotes the total number of classes, TP_c represents the number of true positives for class c, and FN_c is the number of false negatives for class c. In binary settings, this reduces to $BA = (\text{Sensitivity} + \text{Specificity})/2$ [20]. Unlike conventional accuracy, BA cannot be maximized by a "majority-class" classifier; each class must be detected with equal diligence.

3) Macro-Averaged Recall: It computes recall for each class c and then takes their mean, $Recall_{\text{macro}} = \frac{1}{C}\sum_c Recall_c$ because the minority class receives the same weight as the majority. Macro recall directly reflects a model's capacity to find rare positives—an aspect obscured by micro-averaging [21].

4) Matthews Correlation Coefficient (MCC): For the binary case,

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}, \tag{6}$$

where TP denotes the number of true positives, TN is the number of true negatives, FP represents the number of false positives, and FN is the number of false negatives. MCC returns $+1$ for total agreement (every prediction is correct), 0 for an average random prediction (the classifier behaves like a coin toss), and -1 for total disagreement (every prediction is wrong); it is symmetric with respect to class labels and leverages all four confusion–matrix cells, making it resilient to class skew [22].

5) Geometric Mean (G–Mean): It is the geometric mean of per-class recalls,

$$G\text{-}Mean = \left(\prod_{c=1}^{C} Recall_c \right)^{\frac{1}{C}}, \tag{7}$$

where C denotes the total number of classes and $Recall_c$ is the recall for class c. A single poorly predicted class drives the product towards zero; then G–mean rewards classifiers that keep recall high for every class simultaneously. Empirical studies show that maximizing G–mean is tightly linked to improved minority detection in boosting frameworks [23].

6) Weighted Binary Cross-Entropy (WBCE): It augments the standard cross-entropy loss with class weights w_0, w_1:

$$\mathcal{L}_{\text{WBCE}} = -w_1 \cdot y \cdot \log \hat{p} - w_0 \cdot (1-y) \cdot \log(1-\hat{p}). \tag{8}$$

where y denotes the true label (with $y = 1$ for the positive class and $y = 0$ for the negative class), $\hat{p}$ is the predicted probability of the positive class, w_1 and w_0 are the class weights for the positive and negative classes respectively. A common choice is $w_c \propto \frac{1}{N_c}$,where N_c is the number of samples in class c, giving the minority class a proportionally larger impact on the gradients. The WBCE evolved as a practical solution to class imbalance, being widely supported in ML libraries, facilitating its adoption in various applications [24]. Collectively, these metrics expose different facets of performance and, when reported together, provide a fair and informative picture of a classifier's behavior under class imbalance.

IV. Methodology

In this section, the database used in this study is presented. The techniques employed for feature extraction, data preprocessing, and processing are also described. All experiments were conducted on a Lenovo Loq notebook equipped with an Intel Core i5-12450H processor (12th generation).

A. Data

The data for this study were obtained from the George B. Moody PhysioNet Challenge 2025 [25], which combines two databases:

- CODE-15%: it contains over 300,000 12-lead ECG records collected in Brazil between 2010 and 2016. Most recordings have a duration of either 7.3 s or 10.2 s and a sampling frequency of 400 Hz. The binary Chagas labels are self-reported and therefore may or may not have been validated.
- SaMi-Trop: it contains 1,631 12-lead ECG records collected from Chagas patients in Brazil between 2011 and 2012. Most recordings have a duration of either 7.3 s or 10.2 s and a sampling frequency of 400 Hz. The binary Chagas labels are validated by serological tests, and all are positive.

B. Feature Extraction

The raw ECG signals were pre-processed and transformed to generate the features dataset for analysis. The algorithm for ECG feature extraction concerns applying several digital signal processing algorithms for detecting wave amplitudes and characteristic intervals for each of the ECG leads. Considering two demographic features (gender and age) along with all ECG-derived features, a total of 425 features are computed. The first two features correspond to the mean and standard deviation of the ECG samples from lead II. As this lead is particularly important for diagnostic tasks, the other two features, gender and age, are also included to evaluate the relevance of these demographic variables in detecting ChD through ECG analysis.

The first task in ECG signal processing consists of detecting the QRS complex within each lead. For this purpose, an approach that combines Continuous Wavelet Transform (CWT), Hilbert Transform (HT), and a First-Derivative Filter (FDF) is employed [26]. The filtering process is applied over each whole lead signal consisting of a filter cascade composed of CWT, HTs, and FDF. Considering that CWT presents the scale factor as a variable parameter, the approach tests four different scale factors (2^0, 2^1, 2^2, and 2^3).

The detection or adjustment of the most convenient scale factor is intrinsically related to the adaptation of the threshold parameter. Concerning a specific filtered signal version, the approach begins with an initial threshold value equal to 20% of the maximum amplitude within the filtered signal version. If the number of detected QRS occurrences is lower than half the signal interval duration in seconds, the threshold value is decreased by 20%, and the search for QRS fiducial points is repeated. Conversely, if the number of detected QRS fiducial points exceeds twice the signal interval duration in seconds, the threshold is increased by 20%, and the search for QRS fiducial points is repeated.

Once the fiducial points are detected on a given 10-second preprocessed ECG excerpt, an evaluation metric is computed as the product of the standard deviation of the beat-to-beat intervals and the standard deviation of the peak amplitudes. The selected scale factor is that one associated with the filtering process that allows for the detection of fiducial points for which the referred evaluation metric has the lowest value for the test set.

After determining the QRS locations within each lead, the following feature subsets are computed: mean QRS amplitude for each lead (12 features); standard deviation of QRS amplitudes for each lead (12 features); minimum and maximum QRS amplitude for each lead (24 features); mean, standard deviation, minimum, and maximum values of the time differences between the QRS fiducial points of lead II and those of the other leads (48 features); and mean, standard deviation, minimum, and maximum values of the R-R intervals (beat-to-beat intervals) based on the QRS fiducial points of lead II (4 features).

The second task in ECG signal processing consists of QRS complex segmentation. This step employs a method based on the application of the Linear Wavelet Transform (LWT) to the raw ECG signal at scale 2^3, in combination with a non-linear transformation represented by the time-series area curve length (ACL) [27]. The metric ACL is defined by the product between a function related to the area below the waveform and a function representing a curve of the waveform.

A variable threshold is computed throughout the ACL time-series based on the mean and standard deviation of the amplitude values. This variable provides the detection of maximum values, which are associated with R-wave peaks. Here, the R-waves are again used as initial references for searching QRS boundaries. After detecting the maximum critical points associated with QRS fiducial points, a search is performed, starting from the identified peaks and extending to both the right and left sides until local minima of the absolute ACL slope are found. Each local minimum on either side of the R-wave peak reference is associated with a specific QRS boundary.

After determining each QRS boundary location, and consequently, each QRS duration, the following feature subsets are computed: mean, standard deviation, minimum, and maximum values of each QRS area within each lead, where each QRS area is calculated as the sum of squares of samples between the corresponding QRS boundaries (48 features); mean, standard deviation, minimum, and maximum values of the QRS duration, where median values for each QRS boundary are first computed considering all available leads (4 features); mean, standard deviation, minimum, and maximum values of each QRS angle, where each QRS angle is calculated as the arctangent of the ratio between the QRS area measured using lead AVF and the QRS area measured using lead I (4 features); and mean, standard deviation, minimum, and maximum value of ST-segment elevation $(+)$ or depression $(-)$ are computed as the amplitude difference between QRS offset and QRS onset (48 features).

The third task in ECG signal processing consists of P-wave and T-wave detection and segmentation. For this purpose, a method is employed that obtains five scaled versions corresponding to the Discrete Wavelet Transform (DWT) of the ECG input signal at scales 2^1, 2^2, 2^3, 2^4, and 2^5 [19]. After detecting and delineating each QRS complex using the ACL signal computed within scale 2^3 or 2^4, the method enables the detection and delineation of P and T waves.

For this, a specific scaled version is taken, for example, the scale 2^3, and a new ACL signal is derived. In the next step, a local search for two local maxima is performed within the ACL signal between the locations related to QRS offset and the subsequent QRS onset. Actually, the search interval is divided into two segments, each approximately half the R-R interval. It is expected at least one dominant maximum to be found in each of these intervals: The local maximum close to the left QRS offset is associated with a T-wave peak, and the maximum close to the right QRS onset is associated with a P-wave peak.

Concerning each waveform, other specific search windows

are established for identifying the wave edges. Consequently, the algorithm performs a search for local minima between the preceding QRS onset and T-wave peak for delineating the beginning of the T-wave. Analogous searches are performed in three distinct intervals to delineate the respective waveform boundaries. First, the T-wave end is determined by searching between the T-wave peak and half of the R-R interval. Next, the onset of the P-wave is delineated by searching between half of the R-R interval and the P-wave peak. Finally, the end of the P-wave is determined by searching between the P-wave peak and the onset of the subsequent QRS complex.

After determining the locations of the P-wave and T-wave within each lead, several feature subsets are computed: mean P-wave amplitude and mean T-wave amplitude for each lead (24 features); standard deviation of P-wave amplitudes and T-wave amplitudes for each lead (24 features); minimum and maximum P-wave amplitude, and minimum and maximum T-wave amplitude for each lead (48 features); mean, standard deviation, minimum, and maximum values of each P-wave area and each T-wave area within each lead (96 features); mean, standard deviation, minimum, and maximum values of the P-wave duration and T-wave duration are calculated. For this purpose, the median value for each waveform boundary is first computed considering all the available leads (8 features); mean, standard deviation, minimum, and maximum value of each P-wave angle and each T-wave angle, where each waveform angle is computed as the arctangent of the ratio between the waveform area measured using the lead AVF and the waveform area measured using the lead I (8 features); mean, standard deviation, minimum and maximum value of the PR-interval are computed as the time difference of each QRS onset median position across all leads and each P-wave onset median position across all leads (4 features); mean, standard deviation, minimum and maximum value of the PR-segment are computed as the time difference of each QRS onset median position across all leads and each P-wave offset median position across all leads (4 features); mean, standard deviation, minimum, and maximum value of the QT-interval are computed as the time difference of each T-wave end median position across all leads and each QRS onset median position across all leads (4 features); QTc-interval is computed as the mean QT-interval divided by the square root of the mean R-R interval (1 feature).

C. Preprocessing

After the feature extraction process, the final dataset comprised 283,821 samples. Subsequently, data cleaning was performed by removing 2,358 duplicate samples, resulting in a total of 281,463 unique samples. For the classification task, the dataset also includes a label indicating whether each patient has ChD, and the gender (sex) variable was encoded, with 0 representing male and 1 representing female. The data visualization of the target variable was then initiated, which, as shown in Fig. 1, revealed a significant class imbalance.

This imbalance initially created a misleading impression of satisfactory model performance, as the accuracy metric appeared reasonable.

However, a closer investigation showed a high number of false positives, leading to low precision when the model predicts that a patient has ChD. With regard to the features, a noticeable difference in scale was observed. Most features are on a small scale, with the absolute values of the first and second quartiles not exceeding 5.

Furthermore, 13 features exhibit larger scales, such as age and features related to waveform angles, since angles range from -180 to 180. Although feature scaling is typically addressed through normalization, this step was deemed unnecessary in this study because the model employed is tree-based.

Given the misleading initial impression, a data segmentation strategy was adopted in an attempt to mitigate this issue. The two most interpretable features, gender and age, were selected as the basis for segmentation. Each feature was individually analyzed to investigate its behavior in patients with and without ChD. Through data visualization and the application of Wasserstein distance [28] between the two patient samples, it was observed that the proportion of gender remains virtually unchanged regardless of ChD disease status, as illustrated in Fig. 2 and the small value given by the Wasserstein distance (0.0414). In contrast, Wasserstein's distance related to age was considerably greater (7.3354), but not only that, its probability distribution of age exhibits a more significant difference. This is illustrated in Fig. 3, where the ages were grouped as shown in Table I for easier interpretation.

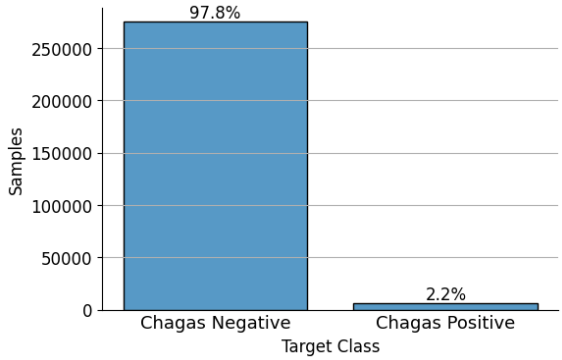


Fig. 1: Distribution of target classes

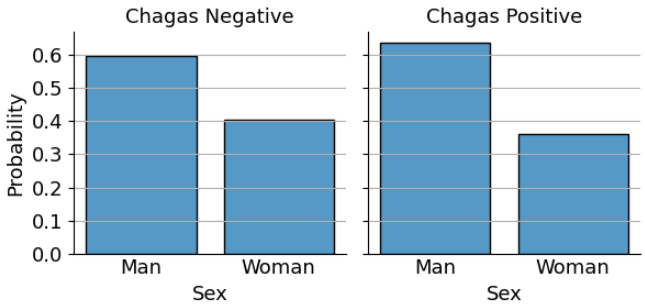


Fig. 2: Distribution of gender across target classes

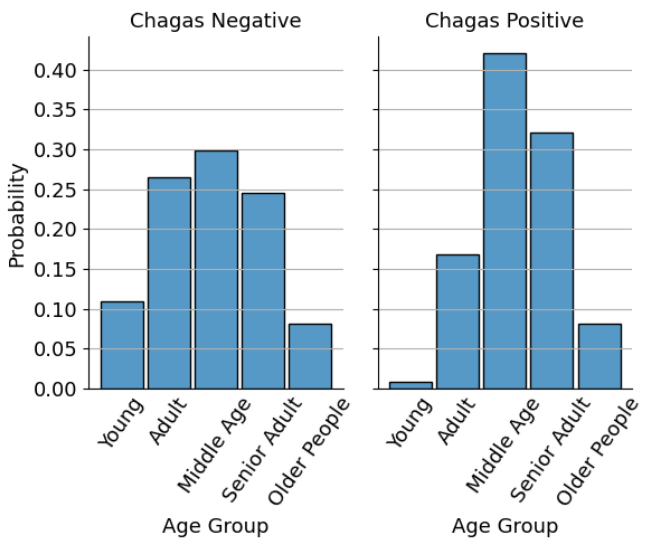


Fig. 3: Distribution of age group across target classes

TABLE I: Age Group Distribution

Age Interval	Age Group Label
(15, 25]	Young
(25, 45]	Adult
(45, 65]	Middle Age
(65, 80]	Senior Adult
(80, 100]	Older People

Based on these observations, the segmentation approach was to create age groups in which the proportion of individuals with ChD is higher than the corresponding proportion among those without ChD. Therefore, as shown in Fig. 3, a very significant difference is observed in the proportion of patients with and without ChD among the middle-aged, senior, and older age groups. Furthermore, the proportion of these age groups is higher among patients with ChD compared to those without ChD. Based on these findings, the age segmentation is defined by patients aged above 45 and those aged 45 or below.

D. Processing and Tuning

The complete dataset was partitioned into two subsets: training (80%) and testing (20%). The CatBoost model was subsequently optimized through the tuning of the following hyperparameters: $iterations$, $depth$, $learning\_rate$, and $l2\_leaf\_reg$. These hyperparameters were optimized through extensive experimentation using 10-fold cross-validation, a procedure adopted to ensure robust performance regardless of data distribution. The optimal values obtained were:

- $iterations$: 720;
- $depth$: 5;
- $learning\_rate$: 0.008;
- $l2\_leaf\_reg$: 0.01.

The WBCE was chosen as the loss function to mitigate the influence of class imbalance. In the next step, the model was evaluated on the test set. As previously discussed, the groups of age ≤ 45 years and > 45 years exhibit markedly different proportions of individuals with and without ChD. Therefore, a CatBoost model was additionally trained and evaluated in each of these age groups. The same hyperparameters identified in the full dataset training were reused. The results of these experiments are presented in the subsequent section.

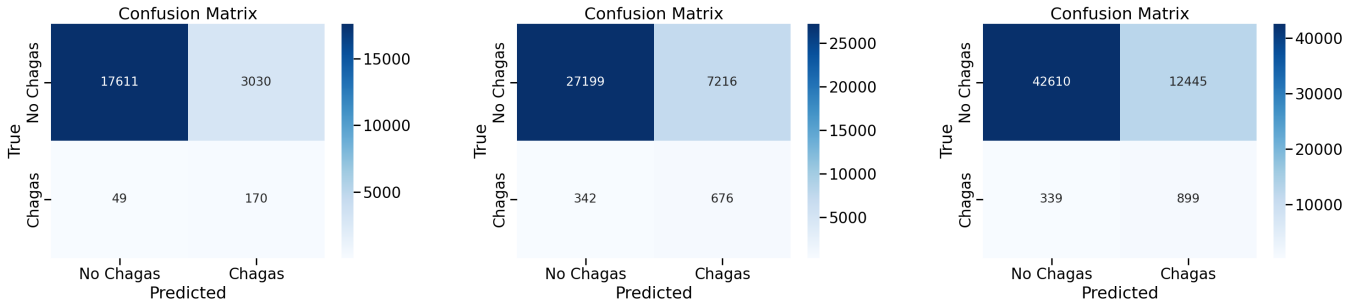
V. Results

Table II presents the mean and standard deviation of each evaluation metric obtained through the cross-validation process, whereas Table III shows the corresponding metric values calculated on the test set. Fig. 4, Fig. 5, and Fig. 6 display, respectively, the confusion matrices, learning curves, and receiver operating characteristic – area under the curve (ROC-AUC) curves for test subsets comprising patients aged 45 years or younger, patients older than 45 years, and the entire test population. Several observations can be made from these results:

- Severe class imbalance: as illustrated in Fig. 4, the proportion of patients with ChD in the test set is approximately $1\% - 3\%$, resulting in an extremely skewed class distribution.
- Optimal fitting: the learning curves suggest that the model achieved optimal fitting because the validation loss closely matches the training loss across all test scenarios, as shown in Fig. 5.
- High stability: throughout cross-validation, the standard deviation of the metrics is consistently small. This indicates that the model is stable across different data splits, and when combined with the optimal fitting, it implies strong generalization capability.
- Age-specific performance gap: performance metrics, both in cross-validation and on the test set, are consistently worse for patients older than 45 years compared to those aged 45 years or younger, indicating that the model has greater difficulty with older individuals. A plausible explanation is the higher prevalence of cardiac abnormalities in older adults; the model may misinterpret some of these abnormalities as markers of ChD.
- Overall metric profile: all metrics exhibit favourable values. The discriminatory power is high, as ROC-AUC always exceeds 0.8, and the calibrated SBS is low (always below 0.2), indicating well-calibrated probability estimates. The only weak metric is the MCC; its low value stems from the extreme class imbalance. Despite the high specificity, the model still yields thousands of false positives (FP) for every hundred true positives (TP).

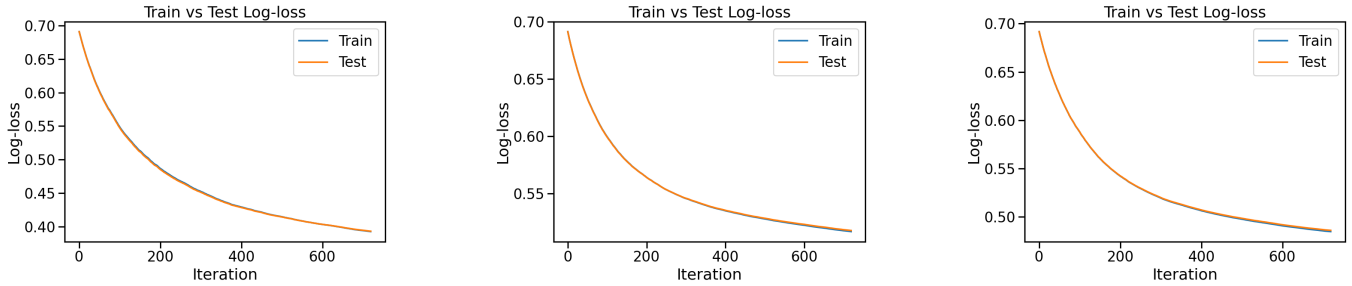
TABLE II: Cross-validation results

Metric	Age ≤ 45	Age > 45	All Ages
ROC-AUC	0.8662 ± 0.0144	0.8083 ± 0.0112	0.8404 ± 0.0065
Balanced Accuracy	0.7800 ± 0.0206	0.7272 ± 0.0131	0.7519 ± 0.0075
Matthews Corr. Coef.	0.1596 ± 0.0110	0.1825 ± 0.0105	0.1750 ± 0.0056
Macro Recall	0.7800 ± 0.0206	0.7272 ± 0.0131	0.7519 ± 0.0075
G-Mean	0.7759 ± 0.0230	0.7243 ± 0.0143	0.7514 ± 0.0077
Stratified Brier	0.1173 ± 0.0029	0.1691 ± 0.0011	0.1569 ± 0.0008
Sensitivity	0.7050 ± 0.0437	0.6641 ± 0.0260	0.7248 ± 0.0135
Specificity	0.8549 ± 0.0065	0.7903 ± 0.0038	0.7790 ± 0.0023



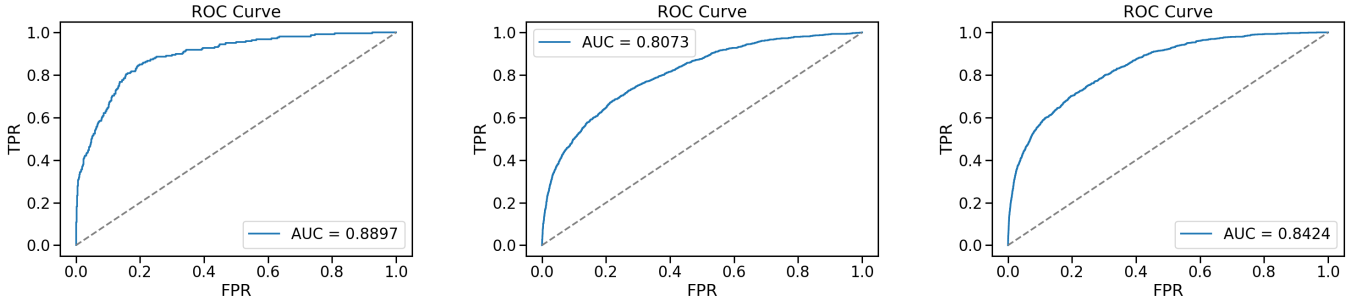
(a) People aged 45 or younger (b) People aged over 45 (c) People of all ages

Fig. 4: Confusion matrices for different age groups.



(a) People aged 45 or younger (b) People aged over 45 (c) People of all ages

Fig. 5: Learning curves (log-loss) for different age groups.



(a) People aged 45 or younger (b) People aged over 45 (c) People of all ages

Fig. 6: ROC curves for different age groups.

TABLE III: Test set results

Metric	**Age** ≤ 45	**Age** > 45	**All Ages**
ROC-AUC	0.8897	0.8073	0.8424
Balanced Accuracy	0.8147	0.7272	0.7501
Matthews Corr. Coef.	0.1780	0.1824	0.1725
Macro Recall	0.8147	0.7272	0.7501
G-Mean	0.8138	0.7244	0.7497
Stratified Brier	0.1180	0.1686	0.1580

VI. Conclusion

This study demonstrates the potential of combining advanced signal processing and ML techniques for the automatic detection of ChD using ECG signals. The CatBoost-based approach yielded robust performance, particularly among younger individuals ($\leq$ 45 years), with consistently high discriminatory power (ROC-AUC $>$ 0.88) and well-calibrated probability estimates. Despite the significant challenge posed by extreme class imbalance, the model maintained stable generalization across data splits and subgroups. Age was confirmed as a key factor for improving classification performance, justifying the chosen segmentation strategy, although the markedly poorer performance in older individuals calls for a more critical examination of its underlying causes (for example, comorbidities, ECG feature overlaps, or data sparsity in that subgroup).

Nonetheless, the results also highlight persistent limitations.

Low precision and F_1-scores, especially for the minority class, underscore the need for further methodological improvements, such as advanced imbalance handling techniques, feature selection, or ensemble methods, to boost sensitivity and specificity for true positive cases. In addition, external validation on independent datasets and integration with other clinical information will be crucial for translating these findings into practice.

The findings reinforce the clinical relevance and feasibility of AI-enhanced ECG analysis for ChD screening in large-scale and resource-limited environments. As AI-driven tools continue to evolve, their adoption could facilitate early intervention and improved outcomes for patients with ChD, supporting public health efforts in endemic and non-endemic regions alike.

ACKNOWLEDGMENT

This work was supported by the National Council for Scientific and Technological Development (CNPq) via grants n°420576/2023-1 and 404683/2024-0.

REFERENCES

[1] K. C. F. Lidani, F. A. Andrade, L. Bavia, F. S. Damasceno, M. H. Beltrame, I. J. Messias-Reason, and M. T. M. Gomes, "Chagas Disease: From Discovery to a Worldwide Health Problem," *Frontiers in Public Health*, vol. 7, p. 166, 2019. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fpubh.2019.00166/full

[2] F. S. Pereira-Silva, M. L. B. C. d. Mello, and T. C. d. Araújo-Jorge, "Doença de Chagas: enfrentando a invisibilidade pela análise de histórias de vida de portadores crônicos," *Ciência & Saúde Coletiva*, vol. 27, no. 5, pp. 1939–1949, 2022. [Online]. Available: https://www.scielo.br/j/csc/a/GzX7xfWzVGt6HXDs78c3qYg/

[3] J. R. Coura, "Chagas disease: what is known and what is needed – A background article," *Memórias do Instituto Oswaldo Cruz*, vol. 102, no. Suppl 1, pp. 113–122, 2007. [Online]. Available: https://www.scielo.br/j/mioc/a/J5P5DTw6G9QmDTDMZLv7jzB/

[4] C. Jidling, D. Gedon, T. B. Schön, C. D. L. Oliveira, C. S. Cardoso, A. M. Ferreira, L. Giatti, S. M. Barreto, E. C. Sabino, A. L. P. Ribeiro, and A. H. Ribeiro, "Screening for Chagas disease from the electrocardiogram using a deep neural network," *PLOS Neglected Tropical Diseases*, vol. 17, no. 7, pp. 1–17, 07 2023. [Online]. Available: https://doi.org/10.1371/journal.pntd.0011118

[5] G. Silva, P. Silva, G. Moreira, V. Freitas, J. Gertrudes, and E. Luz, "A Systematic Review of Ecg Arrhythmia Classification: Adherence to Standards, Fair Evaluation, and Embedded Feasibility," 2025. [Online]. Available: https://arxiv.org/abs/2503.07276

[6] S. A. C. Garzon, A. M. Lorga, and J. C. Nicolau, "Electrocardiography in chagas' heart disease," *São Paulo Medical Journal*, vol. 113, no. 2, pp. 826–834, 1995. [Online]. Available: https://www.scielo.br/j/spmj/a/ScgwKmn8vbhFTmmD9PWRMdm/

[7] H. Narotamo, M. Dias, R. Santos, A. V. Carreiro, H. Gamboa, and M. Silveira, "Deep learning for ecg classification: A comparative study of 1d and 2d representations and multimodal fusion approaches," *Biomedical Signal Processing and Control*, vol. 93, p. 106141, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S174680942400199X

[8] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "Catboost: unbiased boosting with categorical features," *Advances in neural information processing systems*, vol. 31, 2018.

[9] N. Safaei, B. Safaei, S. Seyedekrami, M. Talafidaryani, and A. Masoud, "E-Catboost: An efficient machine learning framework for predicting ICU mortality using the eICU Collaborative Research Database," *PLoS One*, vol. 17, no. 5, p. e0262895, 2022.

[10] Y. Zheng, X. Guo, Y. Yang, and H. Wang, "Phonocardiogram transfer learning-based CatBoost model for diastolic dysfunction identification using multiple domain-specific deep feature fusion," *Computers in Biology and Medicine*, vol. 156, p. 106707, 2023.

[11] C. Li, N. Guan, and H. Zhang, "Anticancer drug synergy prediction based on CatBoost," *PeerJ Computer Science*, vol. 11, p. e2829, 2025.

[12] S. Geeitha, K. Ravishankar, J. Cho, and S. Easwaramoorthy, "Integrating CatBoost algorithm with triangulating feature importance to predict survival outcome in recurrent cervical cancer," *Scientific Reports*, vol. 14, p. 19828, 2024.

[13] D. R. Ghilardi, F. De Rose Ghilardi, G. Silva, T. Vieira, A. Mota, A. L. Bierrenbach, R. F. Damasceno, L. C. Oliveira, and A. D. P. C. Filho, "Machine Learning for predicting Chagas disease infection in rural areas of Brazil," *PLoS Neglected Tropical Diseases*, 2024, accepted 13 March 2024, https://doi.org/10.1371/journal.pntd.0010952.

[14] C. H. L. Cavalcante, P. E. Primo, C. A. F. Sales, W. L. Caldas, J. H. M. Silva, A. H. Souza, E. S. Marinho, R. C. Pedrosa, J. A. L. Marques, H. S. Santos, and J. P. V. Madeiro, "Sudden cardiac death multiparametric classification system for Chagas heart disease patients based on clinical data and 24-hours ECG monitoring," *Mathematical Biosciences and Engineering*, vol. 20, no. 5, pp. 9159–9178, 2023.

[15] A. V. Dorogush, V. Ershov, and A. Gulin, "Catboost: Gradient boosting with categorical features support. arxiv 2018," *arXiv preprint arXiv:1810.11363*, 1810.

[16] C. Bentéjac, A. Csörgő, and G. Martínez-Muñoz, "A comparative analysis of gradient boosting algorithms," *Artificial Intelligence Review*, vol. 54, no. 3, pp. 1937–1967, 2021. [Online]. Available: https://doi.org/10.1007/s10462-020-09896-5

[17] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, no. 9, pp. 1263–1284, 2009.

[18] T. Saito and M. Rehmsmeier, "The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets," *PLoS ONE*, vol. 10, no. 3, p. e0118432, 2015.

[19] G. W. Brier, "Verification of forecasts expressed in terms of probability," *Monthly Weather Review*, vol. 78, no. 1, pp. 1–3, 1950.

[20] K. H. Brodersen, C. S. Ong, K. E. Stephan, and J. M. Buhmann, "The balanced accuracy and its posterior distribution," in *Proceedings of the 20th International Conference on Pattern Recognition*, 2010, pp. 3121–3124.

[21] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Information Processing & Management*, vol. 45, no. 4, pp. 427–437, 2009.

[22] D. Chicco and G. Jurman, "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation," *BMC Genomics*, vol. 21, no. 1, p. 6, 2020.

[23] Y. Sun, M. S. Kamel, A. K. Wong, and Y. Wang, "Cost-sensitive boosting for classification of imbalanced data," *Pattern Recognition*, vol. 40, no. 12, pp. 3358–3378, 2007.

[24] Y. Ho and S. Wookey, "The real-world-weight cross-entropy loss function: Modeling the costs of mislabeling," *IEEE access*, vol. 8, pp. 4806–4813, 2019.

[25] "Detection of Chagas Disease from the ECG: The George B. Moody Physionet Challenge 2025," NIH, 2025.

[26] J. P. Madeiro, P. C. Cortez, J. A. Marques, C. R. Seisdedos, and C. R. Sobrinho, "An innovative approach of QRS segmentation based on first-derivative, Hilbert and Wavelet transforms," *Medical Engineering & Physics*, vol. 34, no. 9, pp. 1236–1246, 2012. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1350453311003250

[27] A. Ghaffari, M. Homaeinezhad, M. Akraminia, M. Atarod, and M. Daevaeiha, "A robust wavelet-based multi-lead electrocardiogram delineation algorithm," *Medical Engineering & Physics*, vol. 31, no. 10, pp. 1219–1227, 2009. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1350453309001647

[28] V. M. Panaretos and Y. Zemel, "Statistical Aspects of Wasserstein Distances," *Annual Review of Statistics and Its Application*, vol. 6, no. 1, p. 405–431, Mar. 2019. [Online]. Available: http://dx.doi.org/10.1146/annurev-statistics-030718-104938