AutoML PyCaret and SHAP explainable AI for ECG signal classification based on amplitude variability

Dmytro Tymoshchuk[1],*, Iryna Didych[1], Andriy Sverstiuk[2], Lyubomyr Mosiy[1] and Yuri Palianytsia[1]

[1] *Ternopil Ivan Puluj National Technical University, Ruska str. 56, Ternopil, 46001, Ukraine*
[2] *I. Horbachevsky Ternopil National Medical University, Maidan Voli St., 1, Ternopil, 46002, Ukraine*

Abstract

Cardiovascular diseases remain the leading cause of global mortality, necessitating advanced non-invasive diagnostic tools. Traditional electrocardiogram (ECG) analysis often focuses on temporal rhythm parameters, frequently overlooking the diagnostic potential of cycle-to-cycle amplitude variability of characteristic waves. In this work, we propose a novel information technology that integrates mathematical modeling of amplitude variability with automated machine learning (AutoML) and explainable artificial intelligence (SHAP). Utilizing open PhysioNet databases, we extracted ten statistical descriptors of amplitude variability to form a dataset comprising four classes: normal, pacemaker, arrhythmias, and morphological abnormalities. The Random Forest model, optimized via the PyCaret library, demonstrated superior performance, achieving an accuracy of over 95% and an Area Under the Curve (AUC) exceeding 0.96 across all classes. Furthermore, SHAP analysis identified Skewness and Kurtosis as the most critical features driving the model's predictions, providing both global and local interpretability. The results confirm that combining amplitude variability descriptors with explainable AutoML frameworks significantly enhances diagnostic precision and transparency, offering a robust foundation for reliable clinical decision support systems.

Keywords

ECG signal, amplitude variability, machine learning, AutoML, PyCaret, explainable AI, SHAP

1. Introduction

Cardiovascular disease (CVD) remains the leading cause of mortality worldwide, accounting for more than 17.9 million deaths annually, according to the World Health Organization (WHO). Electrocardiography, as a non-invasive method of recording the electrical activity of the heart, plays a fundamental role in the early diagnosis and monitoring of cardiac pathologies. The electrocardiogram (ECG) contains critical information about the functional state of the myocardium, encoded in the morphology of the characteristic P, Q, R, S, and T waves and their temporal relationships.

Traditional methods of ECG analysis primarily focus on temporal parameters of heart rhythm (RR intervals, heart rate variability) or morphological features of individual cardiac cycles. However, the cycle-to-cycle amplitude variability of characteristic ECG waves remains insufficiently studied, despite its potential diagnostic value. The amplitude variability function, which reflects the dynamic changes in the amplitude values of the P, Q, R, S, and T waves between successive cardiac cycles, may serve as a sensitive indicator of early pathological alterations in the cardiovascular system (CVS) that precede the clinical manifestation of disease.

Machine learning is already being applied across a wide range of domains, including finance [1], cybersecurity [2, 3], transportation [4], medicine [5], materials science [6, 7], and energy [8]. The integration of machine learning and artificial intelligence methods into cardiac diagnostics opens up unprecedented opportunities for the prediction and early detection of heart disease. Recent studies

ExplAI-2025: Advanced AI in Explainability and Ethics for the Sustainable Development Goals, November 07, 2025, Khmelnytskyi, Ukraine

*Corresponding author.

dmytro.tymoshchuk@gmail.com (D. Tymoshchuk); iryna.didych1101@gmail.com (I. Didych); sverstyuk@tdmu.edu.ua (A. Sverstiuk); lmosiy@gmail.com (L. Mosiy); palanizayb@tntu.edu.ua (Y. Palianytsia)
0000-0003-0246-2236 (D. Tymoshchuk); 0000-0003-2846-6040 (I. Didych); 0000-0001-8644-0776 (A. Sverstiuk); 0009-0000-9778-331X (L. Mosiy); 0000-0002-8710-953X (Y. Palianytsia)

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073
published 2025-12-22

demonstrate that deep learning algorithms achieve high accuracy in ECG classification, significantly outperforming traditional approaches. Advanced AI architectures are capable of automatically detecting complex patterns and hidden regularities in ECG data that often remain invisible during visual interpretation or conventional analysis [9].

The main contribution of this study is the development and experimental validation of an information technology for cardiovascular disease diagnosis based on ECG amplitude variability using AutoML (PyCaret) and Explainable AI (SHAP), which ensures high classification accuracy and interpretability of results.

The remainder of this paper is organized as follows. Section 2 reviews recent research in the field of ECG analysis. Section 3 describes the datasets and the methodology for model development. Section 4 presents the experimental results and discussion. Finally, Section 5 summarizes the conclusions and outlines future research directions.

2. Related works

The current state of research in the field of cardiac signal analysis is characterized by the rapid development of artificial intelligence and machine learning methods for the diagnosis of cardiovascular diseases (CVDs). This systematic review covers key scientific works published between 2024 and 2025, demonstrating a variety of approaches ranging from traditional algorithms to innovative deep learning architectures.

In [10], the authors proposed a new approach to detecting atrial fibrillation and normal sinus rhythm using the concept of TinyML (Embedded Machine Learning). The researchers developed a highly efficient system based on convolutional neural networks, adapted to run on the ESP32 microcontroller. A distinctive feature of their approach is the use of preprocessing of data from the PTB-XL database, including filtering and segmentation of time records into individual heart cycles. The experimental results demonstrate high efficiency: the model achieved an accuracy of 94.1% during training and 94.04% during testing, while the inference accuracy on the microcontroller was 99.33% when using data from a patient simulator. This research opens up new prospects for the creation of portable diagnostic devices with low energy consumption.

Article [11] presents a comprehensive study of the impact of one-dimensional convolutional neural networks on the accuracy of heart rate metrics for electrocardiogram and ballistocardiography (BCG) signals. The researchers focused on the critical problem of motion artifacts, which negatively affect the reliability of vital information such as heart rate. The proposed method for detecting motion artifacts is based on a 1D CNN architecture that analyzes one-second segments of data and classifies them as clean or noisy. The results of the experiments showed a classification accuracy of 95.9% for ECG and 91.1% for BCG signals. The most impressive achievement is the increase in the sensitivity of detection algorithms: from 75% to 98.5% for ECG and from 72.1% to 94.5% for BCG for signals contaminated at 0 dB signal-to-noise ratio.

In their work [12], researchers presented an innovative methodology for predicting heart disease using convolutional neural networks on an expanded PTB-XL+ database. The authors emphasize the importance of automating ECG analysis, as traditional subjective interpretation is labor-intensive and prone to errors. The developed CNN model demonstrates the ability to independently study features from raw data, making it a potentially practical tool for improving diagnostic efficiency. Experimental validation showed an average accuracy of 77.89% in identifying patterns of various heart diseases, including arrhythmias, ischemic heart disease, and myocardial infarction. The results confirm the promise of CNN approaches for improving clinical decision support systems.

Scientists in [13] presented a fundamentally new approach to the classification of cardiac signals by developing an optional multimodal architecture with multiscale receptive fields of a CNN-enhanced transformer. The key innovation is the introduction of switchable modal experts for staged representation: the first stage extracts modality-specific features and balances intermodal relationships, while the second stage captures cross-modal interaction information in a shared latent space. The uniqueness of the architecture lies in its flexibility—thanks to switchable modal experts, the model can be applied to both multimodal and unimodal data. The researchers also solved the problem of performance imbalance between transformers and CNNs by combining the advantages of CNNs to build a CNN-enhanced transformer with improved patch embedding and the integration of convolution and residual connections.

The authors [14] developed an advanced deep learning approach for accurate ECG analysis, which involves both wave delineation and beat-type classification tasks. The researchers integrated two new schemes into the deep learning model. The first scheme represents an adaptive beat segmentation method that determines the optimal duration for each heartbeat based on RR intervals, mitigating segmentation errors from traditional fixed-period segmentation. The second scheme incorporates information about the relative heart rate of the target beat compared to neighboring beats, improving the model's ability to accurately detect premature atrial contractions (PACs). Comprehensive evaluations on the PhysioNet QT, MIT-BIH Arrhythmia, and real-world wearable device datasets demonstrated high performance: 99.81% sensitivity for normal beats, 99.08% for premature ventricular contractions, and 97.83% for PACs. For wave delineation, F1 scores of 0.9842 for non-wave segments, 0.9798 for P waves, 0.9749 for QRS complexes, and 0.9848 for T waves were achieved.

Publication [15] presents a novel machine learning approach for ECG classification using scattering wavelet features. The research methodology includes preprocessing of ECG segments followed by wavelet scattering to extract low-variance features with reduced dimensions. Key features are selected using the Minimum Redundancy and Maximum Relevance (MRMR) algorithm, chosen after a comparative analysis of various feature selection algorithms. The researchers conducted a comprehensive comparative analysis of various machine learning models: Support Vector Machine (SVM), K-Nearest Neighbor (KNN), decision trees, and artificial neural networks with 10-fold cross-validation. Among the twenty models studied, cubic SVM demonstrated the highest accuracy of 99.84%, which indicates the effectiveness of combining wavelet dispersion with optimized machine learning algorithms.

Researchers [16] developed a complex methodology for detecting and classifying heart murmurs based on statistically significant features obtained from comparing spectrogram images of phonocardiogram recordings. The authors used short-time Fourier transform (STFT) to generate spectrograms of PCG signals, which were then compared using Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index Matrix (SSIM). Statistical analysis showed that the SSIM and PSNR similarity indices independently provided 88.23% and 87.94% accuracy, respectively, for distinguishing normal heart sounds from murmurs with a P-value of 2.05×10^{-19}. The best classification results were achieved using a coarse tree with PCA: 85% accuracy during training and 92.50% during testing for the classification of normal heart sounds and different types of murmurs.

Scientists in [17] proposed a comprehensive method for processing cardiac signals that combines wavelet analysis with deep learning algorithms based on artificial intelligence. The scientists used continuous wavelet transforms to calculate scalograms of various cardiac pathologies, creating different types of these signals. The artificial intelligence architecture uses two well-known neural networks, GoogLeNet and SqueezeNet, which have been sufficiently trained in similar applications such as image processing and machine vision. The experimental patient data for the simulation were obtained from the Massachusetts Institute of Technology's standard PhysioNet medical engineering dataset.

In publication [18], the author presented an effective 34-layer ResNet deep network for classifying three types of cardiovascular diseases based on features extracted from the time-frequency domain in the form of scalograms. The researcher combined the proposed ResNet-34 model with transfer learning techniques, demonstrating improved results. The algorithms were compared with other deep networks, such as two different structures of a convolutional neural network (CNN) and a recurrent neural network (RNN), as well as with a classifier based on Sparse Non-Negative Matrix Factorization (SNMF) dictionary learning. The results showed that the ResNet-34-based model has better performance across various evaluation criteria, such as accuracy, sensitivity, and reliability.

Work [19] presented an innovative approach to ECG classification using spiking neural networks (SNNs) with an attention mechanism. The key innovation is the adaptation of trained parameters from artificial neural networks (ANN) to SNN using leaky integrate-and-fire (LIF) neurons. This transfer learning strategy not only leverages the advantages of both types of neural network models but also solves the training problems associated with SNNs. Spiking neural networks, which more accurately mimic brain neural activity through spiking processing, offer a promising path for energy-efficient computing models. Experimental evaluation on two publicly available ECG benchmark datasets showed an overall accuracy of 93.8% on the MIT-BIH Arrhythmia dataset and 85.8% on the PhysioNet Challenge 2017 dataset. These results highlight the potential of SNNs in medical diagnostics, offering a path to more accurate, efficient, and less resource-intensive analyses of heart disease.

In publication [20], researchers proposed a hybrid approach for ECG classification based on machine learning using various digital differentiators and two-dimensional complex wavelet transform (DTCWT). The study presents a systematic approach to classifying ECGs into six different classes based on annotations from the MIT-BIH Arrhythmia database. The methodology includes manual feature extraction using DTCWT to capture critical information from ECGs. Four innovative digital filters are used to differentiate ECGs in order to further enhance the discriminatory power of the extracted features. The Pan-Tompkins algorithm has been improved using these digital differentiators, increasing its effectiveness in detecting QRS complexes.

The paper [21] presents a methodological framework for clustering classification for accurate processing of medical time series. The authors integrated agglomerative hierarchical clustering with representations of Hilbert vector spaces of medical signals and biological sequences. The proposed method demonstrated 96% success in classifying protein sequences by function and effectively identified families in a large set of proteins. In the analysis of cardiac signals, the method retained 0.996 variance in a compressed 6-dimensional space, accurately classifying 87.4% of simulated atrial fibrillation groups and 99.91% of major groups when conduction direction was excluded.

Researchers in [22] developed an innovative approach for classifying heart sounds using harmonic and percussive spectral features from phonocardiograms with a deep feedforward artificial neural network. The methodology includes advanced digital signal processing techniques applied to PCG recordings from the PhysioNet 2016 dataset. A distinctive feature of the approach is the use of harmonic-percussive source separation (HPSS) to extract separate harmonic and percussive spectral features. The feature set consists of 164 attributes, including Chroma STFT, Chroma CENS, mel-frequency cepstral coefficients (MFCC), and statistical features optimized by the ROC-AUC feature selection method. The proposed model achieved a validation accuracy of 93.40% with a sensitivity of 82.40% and a specificity of 80.60%. These results highlight the effectiveness of harmonic features and the reliability of artificial neural networks in classifying heart sounds, especially in resource-constrained environments.

In [23], researchers presented an innovative concept for regenerating cardiac signals using regenerative artificial intelligence. This approach represents a novel use of AI for deep analysis of complex electrical signals generated by the heart. Through advanced AI algorithms, it becomes possible to perform a more in-depth analysis of ECGs, revealing patterns, anomalies, and biomarkers that might otherwise go unnoticed. The goal of cardiac signal regeneration is to identify early signs of heart damage and monitor the heart's ability to regenerate or recover over time. The classification model demonstrates high accuracy for all heartbeat classes, ranging from 93.85% to 99.16%, indicating its effectiveness in detecting and classifying various types of arrhythmias.

Specialists in [24] have developed a three-phase structure for real-time diagnosis of heart disease through behavioral changes in the ECG. The innovative approach first identifies sudden changes in ECG behavior and then determines the cause of these changes through disease classification. The preprocessing stage is integrated with a change point detection (CPD) module, making the structure fully end-to-end and adaptive. The CPD model uses an autoencoder to capture the essential characteristics of ECG in latent space, which are then combined with other temporal features to improve the accuracy of the stack ensemble classifier.

Scientists [25] proposed an alternative approach using end-to-end classification models to remotely obtain a discrete representation of cardiac signals from facial video recordings. Unlike traditional computer vision solutions, which estimate cardiac signals by detecting physical manifestations of heartbeat (such as changes in facial color due to changes in blood oxygenation), the authors introduced a method for discretizing cardiac signals—an innovative preprocessing approach with limited precedents in the health monitoring literature. The results showed that the proposed method outperforms the baseline model on the UBFC-rPPG dataset, reducing the cross-dataset root mean square error from 2.33 to 1.63 beats per minute. Additionally, the approach reduces the computational complexity of post-processing the model output, improving its suitability for real-time applications and deployment on resource-constrained systems.

The aim of our research is to develop information technology for cardiac diagnostics based on the amplitude variability of characteristic ECG waves, using automated machine learning (AutoML) and interpretable artificial intelligence (SHAP) methods to improve the accuracy and reliability of cardiovascular disease diagnosis.

3. Materials and methods

This study used open datasets from PhysioNet. The dataset [26] contained 12-lead electrocardiograms of 45,152 patients recorded at a sampling rate of 500 Hz. This database includes recordings of various common heart rhythm disorders and additional cardiovascular diseases, all of which were annotated by highly qualified experts. The data [27] were generated from a set of long-term ECG recordings of 15 patients (11 men aged 22 to 71 and 4 women aged 54 to 63) with severe congestive heart failure. The dataset [28] contained more than 100 15-minute two-lead ECG recordings. The recordings were accompanied by annotations of the onset, peak, and end of P waves, QRS complexes, T waves, and, where present, U waves for 30–50 selected cardiac cycles in each recording.

To create a dataset for training and testing machine learning models, ECGs were processed using a mathematical model of amplitude variability, which takes into account the amplitude values of characteristic ECG waves (P, QRS, and T) [29]. The resulting dataset contained 651 samples. To build a machine learning model, ten statistical descriptors of ECG amplitude variability were used as input features:

- Mean (arithmetic mean) is a measure of the central tendency of the distribution of amplitude variability;
- Median is a robust characteristic of the central tendency, resistant to the presence of outliers;
- Mode is the most frequently occurring value of amplitude variability;
- Standard Deviation is a measure of dispersion relative to the mean value;
- Sample Variance is the square of the standard deviation, reflecting variability;
- Kurtosis is an indicator of the peakedness (sharpness) of the distribution;
- Skewness is a measure of the asymmetry of the distribution;
- Range is the difference between the maximum and minimum values;
- Minimum is the smallest value of amplitude variability;
- Maximum is the largest value of amplitude variability.

The following symbols were introduced into the dataset: Mean-Mean, Median-Med, Mode-Mo, Standard Deviation-StD, Sample Variance-SV, Kurtosis-Kur, Skewness-Sk, Range-Ra, Minimum-Min, and Maximum-Max. The initial parameter was the diagnosis, from which four classes were formed: class 1 is the conditional norm, class 2 is the conditional norm with an implanted pacemaker, class 3 is arrhythmias, and class 4 is morphological abnormalities. Within the scope of this study, the term morphological abnormalities refers to pathological conditions accompanied by structural changes in the myocardium and/or the cardiac conduction system, which manifest themselves in the form of persistent morphological abnormalities on the ECG. This group includes ischemic heart disease with a history of myocardial infarction (scarring), bundle branch block (complete and incomplete), myocardial hypertrophy, and cardiomyopathies.

To ensure the accuracy of the forecasting quality assessment, the formed set was randomly divided into two unequal parts in a 70/30 ratio, where 70% of the samples were included in the training sample, and the remaining 30% in the test sample. This approach ensured, on the one hand, a sufficient amount of data for effective model training and, on the other hand, created conditions for reliable and objective verification of their predictive ability on new, previously unknown examples.

In this work, the machine learning model was formed using AutoML. The idea behind the AutoML PyCaret method is to create an automated process for building a machine learning model that covers all key stages from initial data preparation to obtaining the final optimized model [30]. The main goal of this approach is to minimize manual intervention by the researcher and significantly accelerate model development while maintaining high accuracy, stability of forecasts, and reproducibility of results. The implementation of PyCaret-type AutoML includes several sequential steps: preliminary data processing, automatic testing of a wide range of machine learning algorithms, evaluation of their effectiveness using a unified set of metrics, and model ranking. Based on this ranking, the best model is automatically selected, which can be further optimized using internal hyperparameter tuning mechanisms (additional training). Thanks to this concept, PyCaret significantly improves the efficiency of the machine learning model creation process, making it transparent, reproducible, and resistant to overfitting. In addition, the use of a single integrated platform simplifies the comparison of different approaches and ensures easy integration into further application systems.

For a deeper understanding of the mechanisms of the constructed model, SHAP analysis (SHapley Additive exPlanations) was applied, which is based on the concept of Shapley values (cooperative game theory) [31]. This method belongs to the class of explainable artificial intelligence (Explainable AI) approaches and allows us to quantitatively assess the contribution of each input feature to the formation of a forecast. At the global level, SHAP allows us to determine the relative importance of features across the entire sample and identify the most influential input features. At the local level, the analysis provides an interpretation of predictions for individual observations, explaining which features and with what intensity influenced the final decision of the model. This two-level interpretation makes it possible to analyze the overall behavior of the algorithm. It also allows for the diagnosis of individual cases to verify the correctness of predictions. The use of SHAP increases the transparency and explainability of the model, which is particularly important in multi-class classification tasks with uneven distribution of samples and complex nonlinear dependencies between features. This helps to increase confidence in the results obtained and allows for more informed decisions based on machine learning model predictions.

4. Results and discussion

In the AutoML system we developed, implemented using the PyCaret library, we built a fully automated process for machine learning model construction that covers all stages — from preliminary data processing to obtaining the final classifier. During the preliminary data processing stage, automatic checks were performed to detect and impute missing values: for numerical features, missing values were replaced using the strategy=‘mean’ principle, and for categorical features, using the strategy=‘most_frequent’ principle. This approach ensured the correctness and consistency of further training. Next, an automatic comparison of a number of classification algorithms was performed, including Random Forest, ExtraTrees, CatBoost, XGBoost, LightGBM, GradientBoosting, multilayer perceptron (MLP), Logistic Regression, Ridge Classifier, and support vector machine (SVM). Each model was evaluated using a unified set of metrics (Accuracy, AUC, Recall, Precision, F1-score, Kappa, MCC), the results of which are shown in Table 1.

Accuracy, recall, specificity, precision, F-score, and G-Mean were calculated using standard methods based on basic classification parameters: true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). For a more comprehensive evaluation of multi-class classification performance, Cohen’s Kappa coefficient of agreement and the Matthews correlation coefficient (MCC) were additionally employed. Both metrics are derived from the complete confusion matrix and enable assessment of prediction balance and reliability even under conditions of class imbalance. Each metric reflects a separate aspect of the model’s performance:

- Accuracy is the total proportion of correctly classified examples;
- Recall measures the model’s ability to correctly identify objects of the target class;

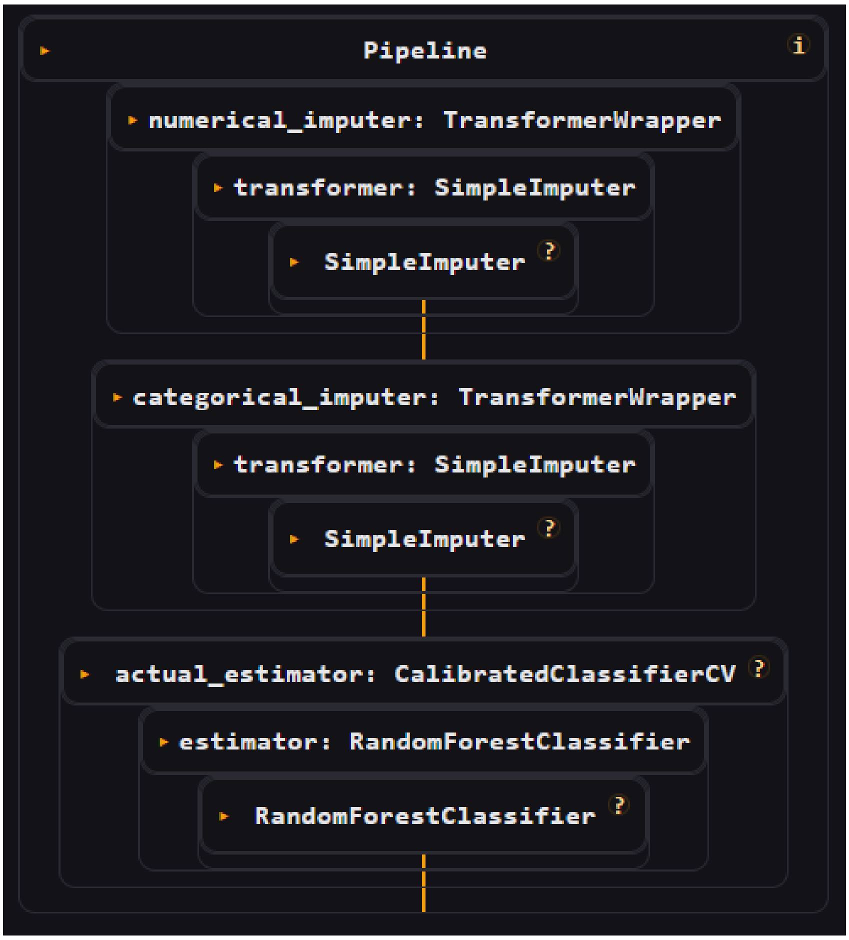
Table 1
Performance evaluation results of machine learning models.

Model	Accuracy	AUC	Recall	Precision	F1	Kappa	MCC
Random Forest	0.9187	0.9433	0.9187	0.9207	0.9186	0.8907	0.8914
Extra Trees	0.9187	0.9402	0.9187	0.9207	0.9186	0.8907	0.8914
CatBoost	0.9165	0.9435	0.9165	0.9186	0.9164	0.8878	0.8886
Extreme Gradient Boosting	0.9121	0.9440	0.9121	0.9143	0.9122	0.8817	0.8824
Light Gradient Boosting	0.9099	0.9438	0.9099	0.9126	0.9099	0.8789	0.8797
Gradient Boosting	0.9033	0.0000	0.9033	0.9063	0.9035	0.8700	0.8709
MLP	0.9011	0.9445	0.9011	0.9036	0.9011	0.8671	0.8679
Logistic Regression	0.7802	0.0000	0.7802	0.7978	0.7792	0.7053	0.7106
SVM	0.7560	0.0000	0.7560	0.8035	0.7453	0.6701	0.6874
Ridge Classifier	0.7055	0.0000	0.7055	0.7072	0.6977	0.6018	0.6063

- Specificity reflects the ability to correctly identify negative cases;
- Precision represents the proportion of correctly predicted positive examples;
- F-score is a balanced assessment that combines Precision and Recall;
- G-Mean is a metric reflecting the balance between Recall and Specificity;
- AUC is the area under the ROC curve;
- Cohen’s Kappa is a coefficient that measures the degree of agreement beyond chance;
- MCC (Matthews Correlation Coefficient) is a robust metric that takes into account all four elements of the confusion matrix and is particularly suitable for imbalanced datasets.

Based on the results of comparison, the Random Forest Classifier model proved to be the most optimal, demonstrating the best balance. To increase the reliability and accuracy of predictions, the selected model was additionally integrated into the probability calibration procedure. To do this, it was wrapped in CalibratedClassifierCV using sigmoid calibration and 5-fold cross-validation (cv=5, ensemble=True). The ensemble=True option means that the model is calibrated on each fold and averages the results, providing more stable probability estimates at the output. This approach improves the correspondence between the predicted probabilities and the actual frequencies of class occurrence, reducing the model’s overconfidence. The random forest is built with hyperparameters selected by AutoML for optimal performance. Specifically, the model contains 240 trees (n_estimators=240) with a maximum tree depth limited to 7 levels (max_depth=7). To prevent overfitting, a restriction is applied to the minimum number of samples in the split node – no less than 10 (min_samples_split=10), and in the leaf – no less than 6 (min_samples_leaf=6). The requirement for a minimum reduction in impurity (impurity criterion) during splitting was set at 0.02 (min_impurity_decrease=0.02), which guarantees that only statistically significant splits are performed. The Gini criterion (criterion=‘gini’) was used to evaluate the quality of the splits. The model did not limit the number of features when selecting splits (max_features=1.0, i.e., all available features at each node are considered), which may increase the completeness of information use. Other parameters remained at their default values: in particular, bootstrap sampling was used to build each tree (bootstrap=True), and a fixed random seed (random_state=42) was used to ensure the reproducibility of the results. The selected configuration — a calibrated Random Forest as part of the PyCaret pipeline — provides not only high classification accuracy but also reliable class probability estimates. This is confirmed by balanced performance indicators for all four classes, which demonstrate the reliability and generalizability of the model. Thus, the model we have built is the result of an automated optimization process and provides high-quality multi-class classification.

Figure 1 shows a diagram of the formed pipeline. The initial stage is numerical_imputer, where the SimpleImputer algorithm can be used for the numerical features Mean, Med, Mo, StD, SV, Kur, Sk, Ra, Min, and Max. The next step is categorical_imputer. The final part of the pipeline is CalibratedClassifierCV, which implements probability calibration for the base RandomForestClassifier algorithm.



Pipeline

numerical_imputer: TransformerWrapper

transformer: SimpleImputer

SimpleImputer

categorical_imputer: TransformerWrapper

transformer: SimpleImputer

SimpleImputer

actual_estimator: CalibratedClassifierCV

estimator: RandomForestClassifier

RandomForestClassifier

Figure 1: Schematic diagram of the pipeline formed by AutoML PyCaret.

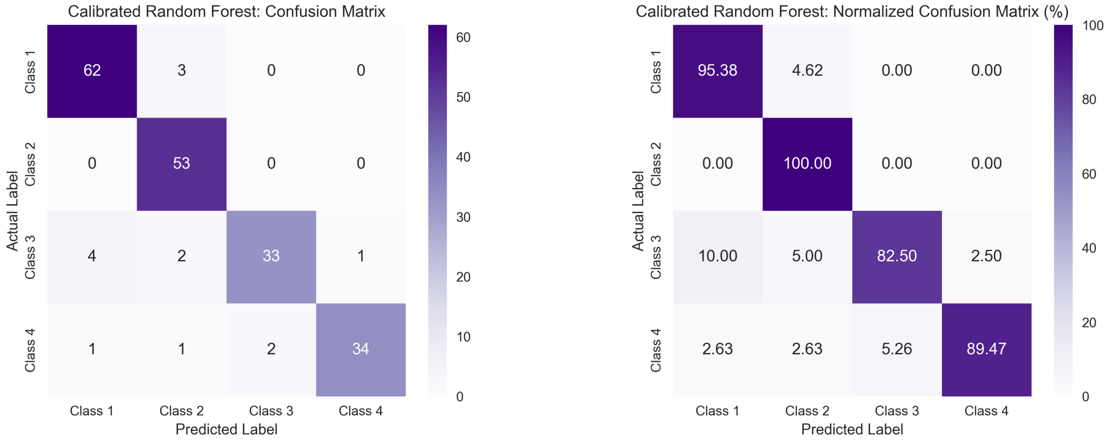
The pipeline shown combines the stages of preliminary data processing and modeling into a single integrated system. This ensures reproducibility, consistency of all steps, and reliability of forecasts, creating a solid foundation for further analysis of classification results.

Figure 2 shows the confusion matrix and its normalized version in percentages (normalized confusion matrix), constructed to evaluate the classification quality of the Random Forest model.

The first matrix shows the absolute values of correctly and incorrectly classified objects in each class, while the second shows relative values in percentage terms. Analysis of the results shows that most objects are correctly assigned to their categories, as evidenced by the high values on the main diagonal. The use of the confusion matrix in two variants, namely absolute and normalized, allows for a comprehensive assessment of the model's performance: on the one hand, in terms of the actual number of correctly and incorrectly classified samples, and on the other hand, taking into account the relative proportions for each class. This allows us to objectively confirm the high efficiency of the constructed model in the task of multi-class classification.

Figure 3 shows the Precision–Recall and ROC curves with the corresponding AUC values for the multi-class classification of the Random Forest model.

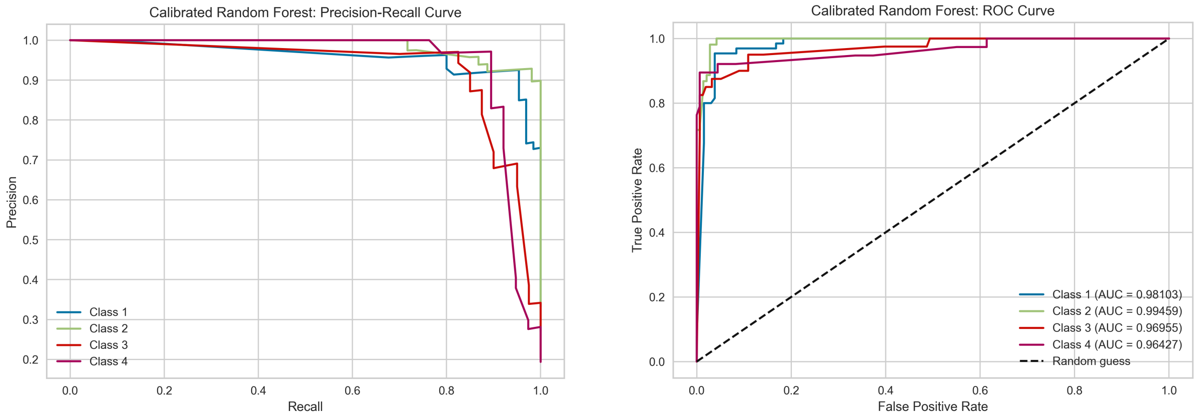
The Precision–Recall curves demonstrate the stability and reliability of the model in the multi-class classification task, confirming its ability to provide high-quality predictions for all four classes. They maintain an accuracy level above 0.9 across a wide range of Recall, indicating the model's effectiveness in detecting objects of each class with a minimum number of false positives. The most consistent results are observed for class 2, which demonstrates the highest stability of indicators. The ROC curves for all four classes are located close to the upper left corner of the graph, confirming the high accuracy of classification. The area under the curve (AUC) exceeds 0.96 in each case, reaching a maximum



(a)

(b)

Figure 2: Confusion matrix (**a**) and normalized confusion matrix (**b**) of the Random Forest model.



(a)

(b)

Figure 3: Precision–Recall (**a**) and ROC curves (**b**) with AUC metrics for multi-class classification.

value of 0.99459 for class 2. This indicates the model's ability to reliably separate positive and negative samples regardless of the selected classification threshold. Thus, the combination of Precision–Recall and ROC analysis results confirms the effectiveness of training and the high generalization ability of the constructed model.

For a more detailed assessment of the classification quality, a set of standard metrics was calculated, the results of which are presented in Table 2.

Table 2
Performance indicators of the Random Forest model.

Class	TP	TN	FP	FN	Accuracy	Recall	Specificity	Precision	F1-Score	G-Mean
1	62	126	5	3	0.9591	0.9538	0.9618	0.9253	0.9393	0.9578
2	53	137	6	0	0.9693	1.0000	0.9580	0.8983	0.9464	0.9787
3	33	154	2	7	0.9540	0.8250	0.9871	0.9428	0.8800	0.9024
4	34	157	1	4	0.9744	0.8947	0.9936	0.9714	0.9315	0.9429

The analysis of the indicators shows that the constructed model provides high-quality multi-class classification. The Accuracy value for all classes exceeds 95%, which confirms the consistency between predictions and actual labels. High Recall and Specificity values demonstrate the model's ability to detect positive and negative examples equally effectively, minimizing the number of false positives and false negatives. The Precision and F1-Score indicators remain consistently high, reflecting the balance between Precision and Recall. Additionally, high G-Mean values indicate the balance of the model's performance in a multi-class task and uneven distribution of examples.

Figure 4 shows the importance rating of input features determined using SHAP analysis for the constructed model. Sk and Kur have the greatest impact on decision-making and are key factors for classification. Min, Ra, and Max are of secondary importance. Other characteristics have a relatively low impact on the model's results.

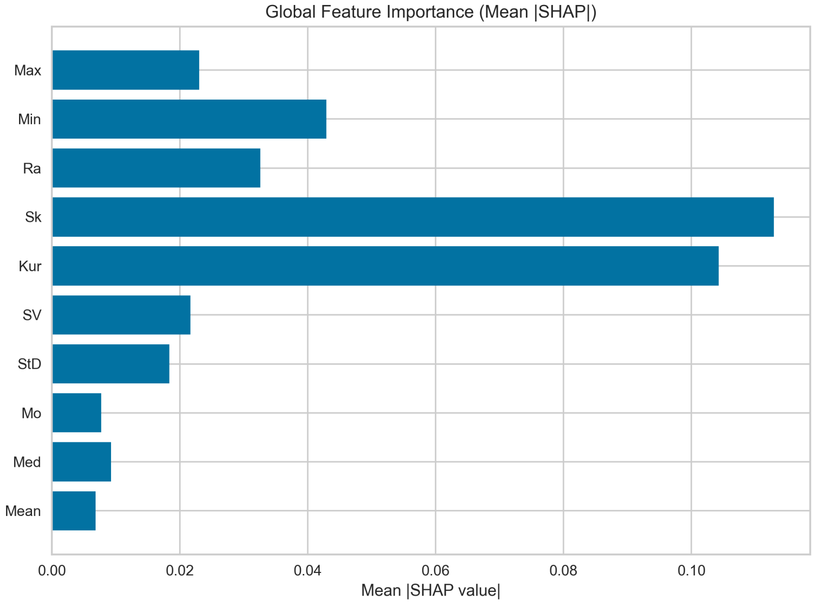


Figure 4: Global Feature Importance based on mean absolute SHAP values.

For a more in-depth interpretation of the classification results, Figure 5 shows SHAP summary diagrams for each of the four classes.

Unlike the global feature importance ranking, which only indicates their relative contribution, these diagrams allow us to assess not only the degree but also the direction of the influence of individual parameters on the probability of belonging to a particular class. The horizontal axis reflects the magnitude of the SHAP value, where positive values indicate an increase in the sample's belonging to the current class, while negative values indicate a decrease in the probability of its classification. Each point corresponds to a separate example from the sample; for better visualization with a large number of observations, the points are partially scattered vertically. The left axis shows the names of the features, sorted by decreasing importance according to the global ranking, and the color scale from blue to red reflects the magnitude of the feature's value (low or high, respectively).

The generalized SHAP graphs for the four classes show clear differences in both the strength of the features' influence and the direction of this influence on the probability of a sample belonging to each class. For class 1, Sk and Kur dominate. High Sk and Kur values can shift the prediction to the right or left, i.e., increase or decrease the probability of class 1, while low Sk values decrease it and low Kur values increase it. Large Min values mainly increase the probability of class 1, while small ones decrease it. For Ra, the opposite picture to Min prevails. Other statistics — SV, Max, StD, Med, Mo, and Mean — make mostly small contributions in terms of modulus. For class 2, Sk is again the leading factor, but the pattern is different: most points with low Sk values are concentrated in the right half-plane, i.e., they increase the probability of belonging to class 2, while an increase in Sk mostly shifts the forecast to the left. At the same time, Kur shows classic monotonicity: higher values increase the probability of

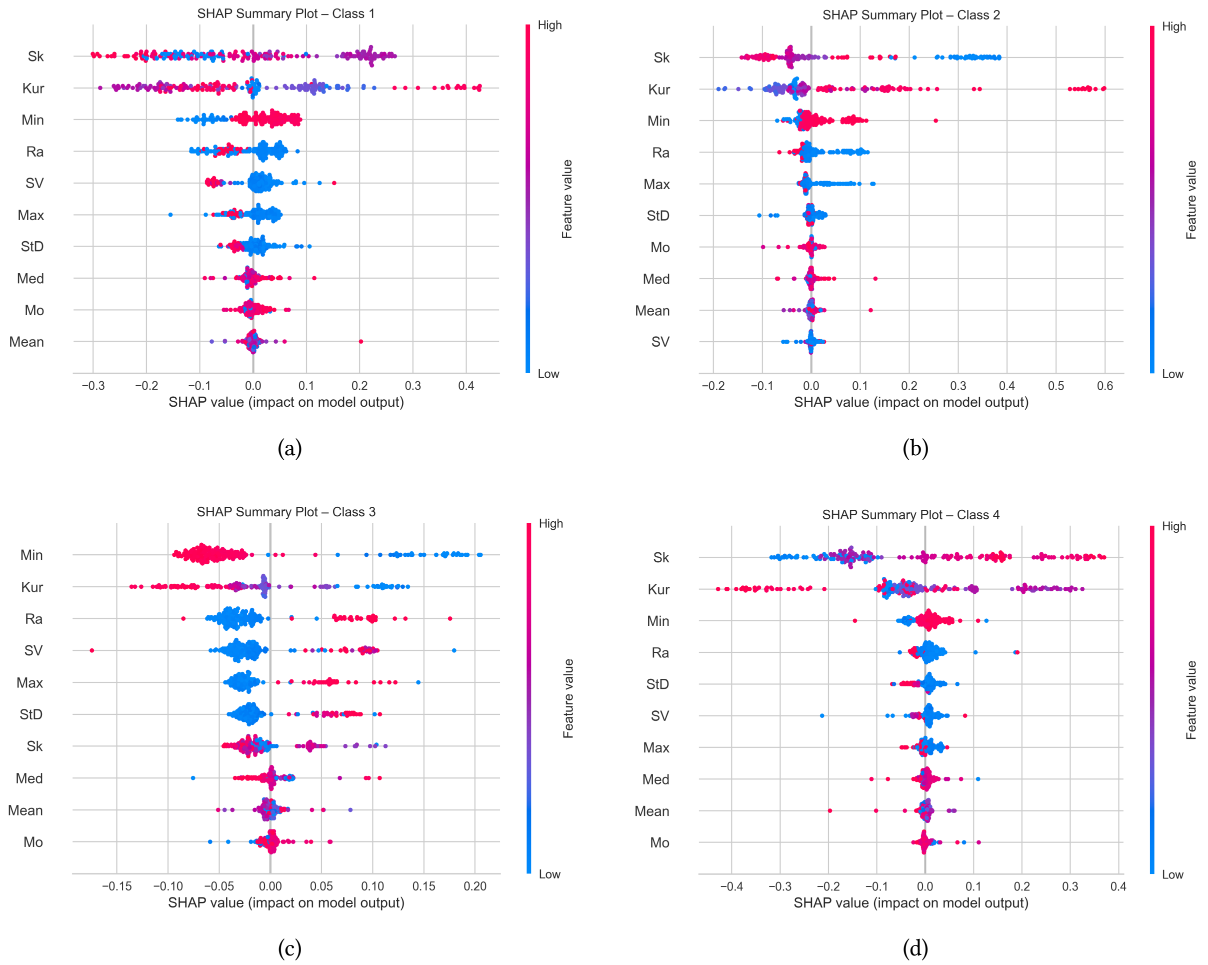
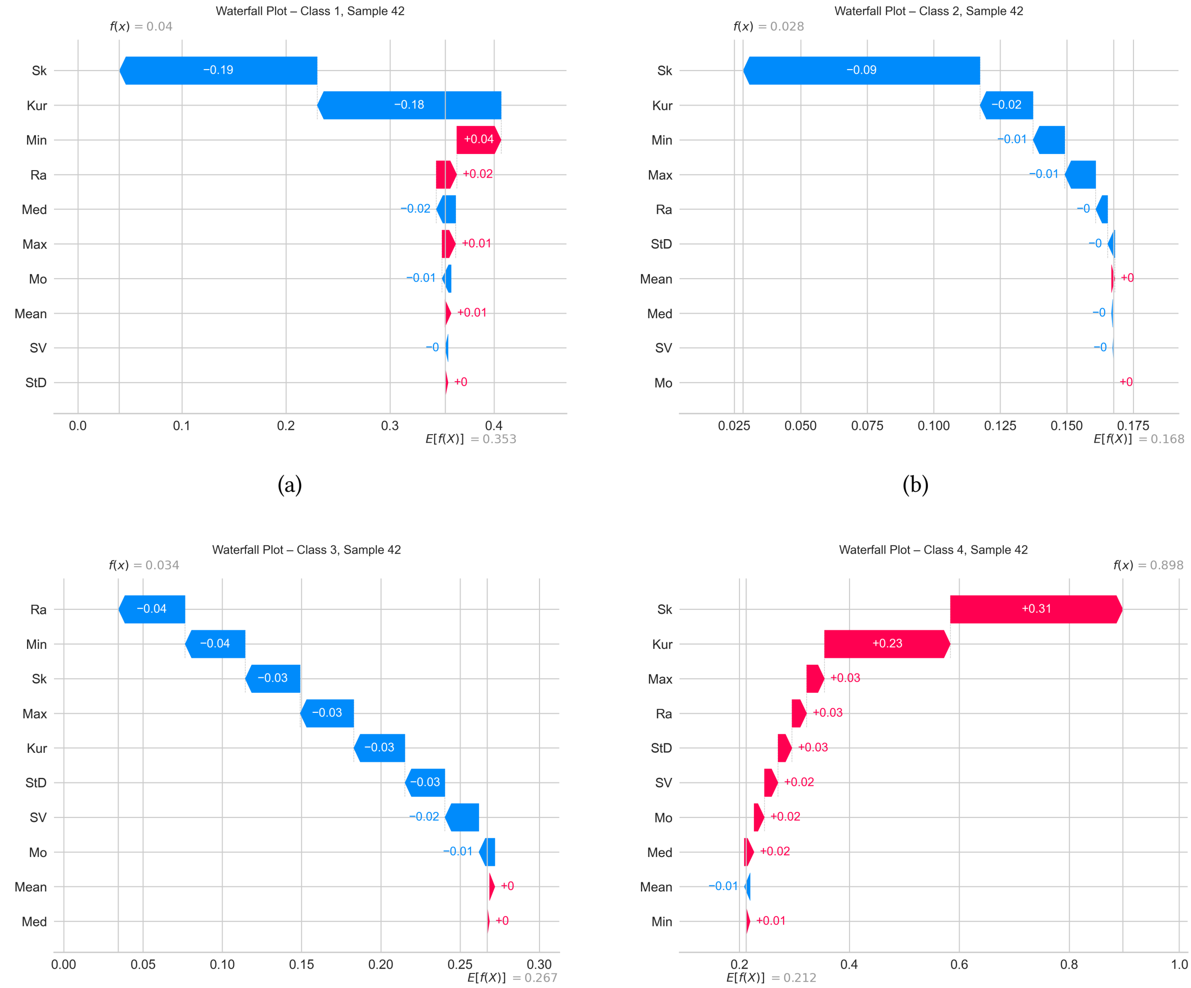


Figure 5: SHAP summary diagrams for four classes: class 1 (**a**), class 2 (**b**), class 3 (**c**), and class 4 (**d**).

class 2, while lower values decrease it. Large Min values increase the probability of class 2, while small values decrease it. Other features provide additional adjustment for class 2 assignment. The profile of class 3 contrasts with the first two. The strongest predictor is Min. Low Min values shift the forecast to the right, increasing the probability of class 3, while high values decrease it. A similar pattern is observed for Kur. Low values of Ra, SV, Max, and StD decrease the probability of class 3, while high values increase it. For the rest of the indicators, the picture is mixed. Class 4 is again dominated by Sk and Kur. High Sk values mainly shift the prediction to the right, i.e., increase the probability of class 4, while low Sk values decrease it. High Kur values can shift the prediction to the right or left, i.e., increase or decrease the probability of class 4, while low values decrease it. Small Min values shift the prediction to the left, decreasing the probability of class 4, while large values mainly increase it. An increase in Ra, StD, SV, and Max indicators is usually accompanied by negative SHAP values, while a decrease is accompanied by positive values. For the rest of the indicators, the picture is mixed.

Thus, the results of the SHAP summary diagrams confirm that the model relies primarily on the statistical characteristics of data distribution (Sk and Kur) in the classification process, and the direction of their influence varies depending on the class. This makes it possible not only to identify key features, but also to understand the specifics of their impact on the formation of predictions, which increases the transparency and explainability of the classification results. At the same time, the patterns obtained reflect only the generalized influence of features, while for individual examples it may differ significantly. Therefore, for a deeper understanding of the model's individual decisions, it is advisable to analyze SHAP waterfall plots, which illustrate the contribution of each feature to the formation of a forecast at the local level.

Figure 6 shows SHAP waterfall diagrams for one sample (Sample 42) in each of the four classes, illustrating the detailed distribution of the contributions of individual features to the final decision of the model.



(a)

(b)

(c)

(d)

Figure 6: SHAP waterfall plots for individual samples (Sample 42) by class: class 1 (**a**), class 2 (**b**), class 3 (**c**), and class 4 (**d**).

Each column reflects the magnitude of the feature's influence: red segments indicate a positive contribution to the probability of belonging to the current class, while blue segments indicate a negative contribution. The SHAP waterfall plots for sample 42 show how the model arrives at an individual prediction $f(x)$ step by step from the base class estimate $E[f(X)]$ thanks to the contributions of individual features. The diagrams clearly show that for class 1 (a), there is a dominant negative influence of Sk and Kur, which significantly reduce the predicted probability. In class 2 (b), the Sk indicator has the greatest negative effect, while the rest of the features have a secondary contribution. For class 3 (c), almost all features have a negative impact on the probability of the selected class. In contrast, class 4 (d) shows the opposite situation. The Sk and Kur indicators provide the most significant positive contribution, significantly shifting the model's prediction in favor of this class. Other features have a mostly positive but insignificant impact. Thus, waterfall diagrams allow us to track not only the relative importance of features, but also the specific direction of their influence on the model's decision for individual examples, providing a deeper interpretation of the classifier's performance at the local level. This ultimately leads to the sample being assigned to this class with a probability of 0.898.

The use of SHAP analysis provided a comprehensive understanding of the model's mechanisms by combining global and local interpretations of its decisions. SHAP analysis ensured the transparency and explainability of the classification results, allowing us to identify key predictors and assess their role in decision-making. The results obtained further confirm the sensitivity of the amplitude variability function to changes in the state of the human cardiovascular system and allow, on its basis, the use of additional informative features in the form of the above statistical estimates for cardiac diagnostics.

The proposed machine learning method has formed the basis for the development of an information technology for cardiological diagnostics. However, it should be noted that the diagnostic specificity of this technology is limited, since not all cardiovascular pathologies manifest through changes in the amplitude of ECG waves. In particular, certain rhythm disturbances, conduction abnormalities, or ischemic conditions may have only a minimal effect on amplitude variability, which restricts the universality of the method.

5. Conclusion

The study developed and experimentally tested an information technology for automated diagnosis of cardiovascular pathologies based on the amplitude variability of ECG waves, combined with AutoML (PyCaret) and explainable artificial intelligence (SHAP). The use of AutoML ensured an efficient and reproducible model development process that minimizes manual intervention by the researcher. Among the tested algorithms, Random Forest with probability calibration indicates the highest performance, achieving high Accuracy (over 95%) and AUC (over 0.96) values for all classes. This demonstrates the model's ability to reliably distinguish between normal, pacemaker patients, arrhythmias, and morphological abnormalities. SHAP analysis enabled the identification of the key features that most influence the model's decisions and allowed tracking their impact at both global and local levels. Specifically, Skewness (Sk) and Kurtosis (Kur) were identified as the dominant predictors affecting classification probability. This interpretation increased the transparency and trust in predictions, which is especially important for medical applications. The results confirm the feasibility of integrating AutoML and explainable AI into cardiac diagnostics, opening up prospects for the creation of reliable, interpretable, and practically applicable clinical decision support systems. This study has a limitation, as the proposed machine learning–based diagnostic technology relies on ECG amplitude variability, which does not capture all cardiovascular pathologies.

In subsequent studies, appropriate ML methods will be applied, and SHAP analysis will be performed for the time variability function, taking into account the amplitudes of characteristic ECG waves.

Acknowledgments

The authors would like to thank the ERASMUS-EDU-2025-CBHE project "Digital Transformation and Curriculum Development for Healthcare Teams" (Digi-CHange - 101233888 – GAP-101233888) for the idea, inspiration and funds that made this work possible.

Declaration on Generative AI

During the preparation of this work, the authors used Grammarly in order to grammar and spell check, and improve the text readability. After using the tool, the authors reviewed and edited the content as needed to take full responsibility for the publication's content.

References

[1] S. Levantesi, G. Zacchia, Machine learning and financial literacy: An exploration of factors influencing financial knowledge in italy, Journal of Risk and Financial Management 14 (2021) 120. doi:10.3390/jrfm14030120.

[2] D. Tymoshchuk, O. Yasniy, M. Mytnyk, N. Zagorodna, V. Tymoshchuk, Detection and classification of ddos flooding attacks by machine learning method, in: CEUR Workshop Proceedings, volume 3842, 2024, pp. 184–195.

[3] Y. Klots, V. Titova, N. Petliak, V. Cheshun, A.-B. Salem, Research of the neural network module for detecting anomalies in network traffic, in: CEUR Workshop Proceedings, volume 3156, 2022, pp. 378–389.

[4] U. K. Lilhore, A. L. Imoize, C.-T. Li, S. Simaiya, S. K. Pani, N. Goyal, A. Kumar, C.-C. Lee, Design and implementation of an ml and iot based adaptive traffic-management system for smart cities, Sensors 22 (2022) 2908. doi:10.3390/s22082908.

[5] S. O. Nykytyuk, A. S. Sverstiuk, S. I. Klymnyuk, D. S. Pyvovarchuk, Y. B. Palaniza, Approach to prediction and receiver operating characteristic analysis of a regression model for assessing the severity of the course lyme borreliosis in children, Rheumatology 61 (2023) 345–352. doi:10.5114/reum/173115.

[6] D. Tymoshchuk, O. Yasniy, P. Maruschak, V. Iasnii, I. Didych, Loading frequency classification in shape memory alloys: A machine learning approach, Computers 13 (2024) 339. doi:10.3390/computers13120339.

[7] O. Yasniy, D. Tymoshchuk, I. Didych, V. Iasnii, I. Pasternak, Modelling the properties of shape memory alloys using machine learning methods, Procedia Structural Integrity 68 (2025) 132–138. doi:10.1016/j.prostr.2025.06.033.

[8] W. Pannakkong, V. T. Vinh, N. N. M. Tuyen, J. Buddhakulsomsiri, A reinforcement learning approach for ensemble machine learning models in peak electricity forecasting, Energies 16 (2023) 5099. doi:10.3390/en16135099.

[9] O. Kovalchuk, O. Barmak, P. Radiuk, L. Klymenko, I. Krak, Towards transparent AI in medicine: ECG-based arrhythmia detection with explainable deep learning, Technologies 13 (2025) 34. doi:10.3390/technologies13010034.

[10] G. V. B. F. Silva, M. D. Lima, J. A. F. Filho, M. J. Rovai, Atrial fibrillation and sinus rhythm detection using tinyml (embedded machine learning), in: IFMBE Proceedings, volume 104, Springer Nature Switzerland, Cham, 2024, pp. 633–644. doi:10.1007/978-3-031-49407-9_63.

[11] J. P. Moreno, M. A. Sepúlveda, E. J. Pino, 1d convolutional neural network impact on heart rate metrics for ecg and bcg signals, Journal of Medical and Biological Engineering (2024). doi:10.1007/s40846-024-00872-w.

[12] T. Soni, D. Gupta, M. Uppal, Deciphering cardiac signals: Leveraging cnns on the ptb-xl+ ecg database, in: 2024 Second International Conference on Intelligent Cyber Physical Systems and Internet of Things (ICoICI), IEEE, 2024, pp. 1–5. doi:10.1109/icoici62503.2024.10696119.

[13] T. Zhang, C. Lian, B. Xu, Y. Su, Z. Zeng, Cardiac signals classification via optional multimodal multiscale receptive fields cnn-enhanced transformer, Knowledge-Based Systems 300 (2024) 112175. doi:10.1016/j.knosys.2024.112175.

[14] J. Lim, D. Han, M. Pirayesh Shirazi Nejad, K. H. Chon, Ecg classification via integration of adaptive beat segmentation and relative heart rate with deep learning networks, Computers in Biology and Medicine 181 (2024) 109062. doi:10.1016/j.compbiomed.2024.109062.

[15] S. K. Janani, R. Sabeenian, Machine learning-based ecg classification using wavelet scattered features, AIUB Journal of Science and Engineering (AJSE) 23 (2024) 168–176. doi:10.53799/ajse.v23i2.821.

[16] P. Careena, M. M. S. J. Preetha, P. Arun, Statistically significant feature-based heart murmur detection and classification using spectrogram image comparison of phonocardiogram records with machine learning techniques, Australian Journal of Electrical and Electronics Engineering (2024) 1–15. doi:10.1080/1448837x.2024.2312491.

[17] M. Zilaie, Z. Mohammadkhani, K. A. Asrari, S. R. Talebiyan, Heart signal processing using wavelet analysis and deep learning algorithms based on artificial intelligence, in: Lecture Notes in Electrical Engineering, volume 1234, Springer Nature Switzerland, Cham, 2025, pp. 571–584. doi:10.1007/978-3-031-84100-2_68.

[18] S. Mavaddati, Ecg arrhythmias classification based on deep learning methods and transfer learning

technique, Biomedical Signal Processing and Control 101 (2025) 107236. doi:10.1016/j.bspc.2024.107236.

[19] A. Rana, K. K. Kim, Electrocardiography classification with leaky integrate-and-fire neurons in an artificial neural network-inspired spiking neural network framework, Sensors 24 (2024) 3426. doi:10.3390/s24113426.

[20] H. K. P. Katamreddi, T. K. Battula, A hybrid approach for machine learning based beat classification of ecg using different digital differentiators and dtcwt, Computers in Biology and Medicine 194 (2025) 110426. doi:10.1016/j.compbiomed.2025.110426.

[21] S. Ruipérez-Campillo, M. Reiss, E. Ramírez, A. Cebrián, J. Millet, F. Castells, Clustering and machine learning framework for medical time series classification, Biocybernetics and Biomedical Engineering 44 (2024) 521–533. doi:10.1016/j.bbe.2024.07.005.

[22] A. Singh, V. Arora, M. Singh, Heart sound classification using harmonic and percussive spectral features from phonocardiograms with a deep ann approach, Applied Sciences 14 (2024) 10201. doi:10.3390/app142210201.

[23] A. Mohaamed Salman, T. Sivasakthi, S. Brindha, M. Suresh kumar, Cardiac signal regeneration using regenerative ai, in: 2025 International Conference on Computing and Communication Technologies (ICCCT), IEEE, 2025, pp. 1–6. doi:10.1109/iccct63501.2025.11019936.

[24] M. Gupta, R. Wadhvani, A. Rasool, Deep learning-based real-time diagnosis of cardiac diseases through behavioral changes in ecg signals, Biomedical Signal Processing and Control 104 (2025) 107532. doi:10.1016/j.bspc.2025.107532.

[25] M. Moustafa, M. A. Farooq, A. Elrasad, J. Lemley, P. Corcoran, Visual cardiac signal classifiers: A deep learning classification approach for heart signal estimation from video, IEEE Access (2024). doi:10.1109/access.2024.3472508.

[26] J. Zheng, H. Guo, H. Chu, A large scale 12-lead electrocardiogram database for arrhythmia study (version 1.0.0), PhysioNet, 2022. doi:10.13026/WGEX-ER52.

[27] D. S. Baim, W. S. Colucci, E. S. Monrad, H. S. Smith, R. F. Wright, A. Lanoue, D. F. Gauthier, B. J. Ransil, W. Grossman, E. Braunwald, Survival of patients with severe congestive heart failure treated with oral milrinone, Journal of the American College of Cardiology 7 (1986) 661–670. doi:10.1016/s0735-1097(86)80478-8.

[28] P. Laguna, R. G. Mark, A. Goldberg, G. B. Moody, A database for evaluation of algorithms for measurement of qt and other waveform intervals in the ecg, in: Computers in Cardiology 1997, IEEE, 1997, pp. 673–676. doi:10.1109/cic.1997.648140.

[29] A. Sverstiuk, L. Mosiy, Mathematical modeling of electrocardiogram signal amplitude variability for information technology analysis of their morphological and rhythmic characteristics, Computer-Integrated Technologies: Education, Science, Production 59 (2025) 228–240. doi:10.36910/6775-2524-0560-2025-59-29.

[30] Dasarpai, automl-pycaret: An open-source, low-code machine learning library in python, 2024. URL: https://github.com/dasarpai/automl-pycaret.

[31] S. Lundberg, shap: A game theoretic approach to explain the output of any machine learning model, 2025. URL: https://github.com/shap/shap.