Сучасний стан наукових досліджень та технологій в промисловості. 2026. № 1 (35) *ISSN 2522-9818 (print)*
Innovative technologies and scientific solutions for industries. 2026. No. 1 (35) *ISSN 2524-2296 (online)*

UDC 004.75:37.018.43 DOI: https://doi.org/10.30837/2522-9818.2026.1.055

Alexander Krajči, Ludmila Sidorenko, Olesia Barkovska

PREDICTING RISKS OF CARDIOVASCULAR DISEASE ON SMALL DATASETS USING FEATURE ENGINEERING

Relevance. Cardiovascular diseases remain a leading cause of mortality globally, creating a high demand for automated diagnostic systems. However, developing reliable machine learning models for electrocardiogram (ECG) analysis is often hindered by the availability of only small-scale and imbalanced datasets, which limits the effectiveness of deep learning approaches. **The object of research** is the process of automated processing and classification of electrocardiographic signals for diagnostic purposes. **The subject of the research** includes methods of beat-centric feature extraction, patient-level aggregation strategies, and machine learning algorithms for cardiovascular risk prediction. **The purpose of this paper** is to develop and evaluate a reliable classification framework, optimized for small datasets, that increases prediction accuracy by leveraging patient-level feature aggregation and explainable machine learning models. To achieve this goal, the following tasks were solved: 1) implementation of a robust preprocessing pipeline using a refined Pan-Tompkins algorithm for precise beat-centric segmentation; 2) development of a statistical feature aggregation strategy to mitigate local signal variability; and 3) optimization and validation of a Random Forest classifier. The methodology employed includes digital signal processing (Butterworth filtering), advanced feature engineering (HRV, Wavelets analysis), and rigorous 10-fold Stratified Cross-Validation to ensure generalization on limited data. **Research results.** The study proposes a pipeline initiating with standard signal preprocessing, followed by precise R-peak detection and beat-centric segmentation. Physiological features (HRV, wavelet, morphological) are then extracted from individual segments and statistically aggregated at the patient level. Experiments on a dataset of 164 subjects demonstrated that the proposed patient-level aggregation strategy significantly outperformed traditional segment-based analysis. The final Random Forest model achieved an ROC-AUC score of 0.84. Feature importance analysis confirmed the critical role of Heart Rate Variability (HRV) metrics, particularly SDNN and RMSSD, in differentiating between healthy and high-risk subjects.

Keywords: classification; risk prediction; feature importance; machine learning; ECG signal; Random Forest.

Introduction

Relevance

Cardiovascular diseases (CVDs) remain the leading cause of global mortality, driving the urgent demand for automated electrocardiogram (ECG) analysis systems [1]. In practice, however, model development is frequently constrained by small, imbalanced, and partially noisy datasets, a setting in which end-to-end deep learning often underperforms or becomes fragile without extensive data curation and augmentation. Consequently, feature-engineering-driven pipelines combined with interpretable machine learning remain highly relevant for clinical screening scenarios and low-data environments [2].

A persistent methodological gap arises from how ECG segments are handled. Many studies classify unaligned fixed-length windows, implicitly assuming that a random window is representative of a patient's global cardiac status. In small datasets, this assumption is brittle: local artifacts within a single window can dominate predictions, and segment-level labels may leak patient-specific idiosyncrasies across folds unless evaluation is performed strictly at the patient level. Addressing these issues, this work adopts a beat-centric segmentation anchored at R peaks, aggregates features at the patient level, and trains an explainable ensemble model to improve generalization in small cohorts.

An overview of scientific works

Many studies focus on automated ECG signal processing and CVD detection. Classical surveys summarize the landscape of ECG analysis, covering preprocessing, morphological characterization, and feature extraction strategies for arrhythmia and risk stratification tasks (e.g., statistical and spectral descriptors, wavelets, entropy-based measures) [3, 4, 5]. Deep Learning approaches in cardiovascular diagnostics are reviewed in [6], which highlights both their potential and the difficulties posed by data needs. Foundational signal processing elements include QRS detection (e.g., Pan–Tompkins) and more recent advances in precise R peak localization, which underpin robust beat-centric workflows [7–10]. Studies [7, 11] focus on current trends in feature extraction methods and their applications. Higher-order wavelet packet decomposition (WPD) statistics are considered in [8]. In terms of signal preprocessing, [9]

ISSN 2522-9818 (print) Сучасний стан наукових досліджень та технологій в промисловості. 2026. № 1 (35)
ISSN 2524-2296 (online) Innovative technologies and scientific solutions for industries. 2026. No. 1 (35)

establishes the basic real-time QRS detection algorithm, and [10] presents more recent developments in accurate R-peak localization. In addition to preprocessing, classifiers are important in processing ECG signals [12, 13]. In parallel with signal processing, the interpretability of decision systems has motivated the adoption of tree-based models and fuzzy inference [14, 15]. In particular, fuzzy decision trees and fuzzy classifiers have been shown to provide transparent rule-based reasoning and to support importance analysis of decision-making factors, which is vital for safety-critical healthcare applications [13, 16]. These efforts range from early formulations of fuzzy decision tree induction and information measures, to reliability assessment of healthcare systems via fuzzy trees, and to modern, large scale ECG classification with fuzzy approaches. Collectively, this body of work motivates our preference for interpretable, feature-based modeling in low data regimes.

Problem Statement and Approach

We target cardiovascular risk prediction on a small proprietary cohort of 164 subjects (92 high risk, 72 healthy), recorded at 512 Hz, reflecting a mild class imbalance representative of real-world settings. Rather than classifying unaligned windows, we first perform standard artifact suppression, precise R peak detection, and beat centric segmentation; then we extract a comprehensive set of physiologically grounded features (time/frequency domains, HRV metrics, wavelet energy, morphology, non-linear entropy). To stabilize subject profiles and mitigate local variability, we aggregate segment-level features per patient via summary statistics (mean, standard deviation, min, max), and train an interpretable Random Forest with systematic feature selection and cross-validated hyperparameter tuning. All validation is carried out with stratified 10-fold cross validation at the patient level to prevent leakage; no synthetic oversampling (e.g., SMOTE) is used in final evaluation, ensuring performance reflects the natural class distribution.

Our emphasis on interpretable features and transparent model behavior is consistent with the broader trend of leveraging fuzzy decision trees and fuzzy classifiers for explainability in medical AI. Fuzzy trees support rule extraction and factor importance analysis, which can complement or extend tree ensemble importance measures and SHAP style attributions. Recent research demonstrates that fuzzy frameworks can successfully handle uncertainty, vagueness, and limited data, especially in clinical contexts. We view our pipeline as a strong baseline for small-cohort ECG studies, which can be further hybridized with fuzzy decision tree induction to yield explicit clinical rules while retaining the robustness of patient-level aggregation [13, 14].

Setting objectives

A key difficulty in implementing machine learning on small ECG groups is the approach for managing segmented data. Classifying unaligned segments independently often introduces variability, as local artifacts in a single segment may not reflect the patient's overall cardiac status. Therefore, it becomes necessary to optimize the processing pipeline by implementing a robust patient-level aggregation strategy. Based on this, the goal of this paper is to develop and evaluate a reliable classification framework, optimized for small datasets, that increases prediction accuracy by leveraging patient-level feature aggregation and explainable machine learning models. This paper makes the following contributions to ECG based risk prediction under small data constraints:

1. ***Beat-centric, patient-level pipeline for small cohorts***. We formalize and evaluate an end-to-end workflow that combines zero-phase bandpass filtering, refined Pan–Tompkins–inspired R-peak detection, beat-anchored segmentation, and patient-level feature aggregation. The design explicitly targets low sample settings where segment-only classification is unstable.

2. ***Patient level aggregation as a principled regularizer*.** We introduce a simple yet effective statistical aggregation schema (mean, std, min, max) over segment features per subject and demonstrate its superiority to segment-based analysis on the same data, improving ROC AUC from 0.64 (baseline) to 0.84 and raising accuracy from 0.57 to 0.80. This quantifies the benefit of aggregation for reducing local noise and enhancing generalization in small populations.

3. ***Rigorous, leakage free evaluation.*** All experiments use stratified 10-fold cross validation at the patient level, with no oversampling in final reporting, and a feature selection step driven by Random Forest importance to curb dimensionality and improve stability – practices aligned with clinical ML standards for limited datasets.

4. ***Clinically meaningful interpretability***. We provide feature importance analysis showing that HRV metrics (e.g., SDNN, RMSSD) dominate the model's decisions, in agreement with physiological understanding of autonomic regulation – thereby enhancing trust and facilitating clinical translation.

Сучасний стан наукових досліджень та технологій в промисловості. 2026. № 1 (35) *ISSN 2522-9818 (print)*
Innovative technologies and scientific solutions for industries. 2026. No. 1 (35) *ISSN 2524-2296 (online)*

5. ***Pathway to fuzzy explainable hybrids.*** By situating our results within the literature on fuzzy decision trees and fuzzy classifiers, we outline a natural extension toward rule based, linguistically interpretable models on top of patient level features, potentially combining the robustness of ensembles with the transparency of fuzzy rules.

Materials and Methods

ECG Signal Morphology

Before discussing the dataset, it is essential to define the structural characteristics of the analyzed signals. A typical ECG cycle (Fig. 1) consists of distinct waves reflecting the electrical activity of the heart: the P-wave (atrial depolarization), the QRS complex (ventricular depolarization), and the T-wave (ventricular repolarization). In the context of cardiovascular risk prediction, the QRS complex is particularly significant as it represents the main contraction of heart ventricles. Its accurate detection is a prerequisite for extracting heart rate variability (HRV) metrics, which are derived from the R-R intervals (the time distance between consecutive R-peaks). This morphological structure is susceptible to distortion and noise, which might result in inaccurate feature extraction and call for thorough preprocessing.

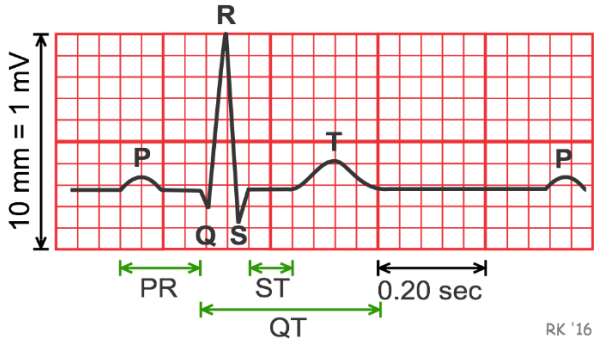

Fig. 1. Morphology of a standard ECG cycle composed of P,Q,R,S and T waves [17]

Dataset description

The experimental evaluation was conducted using a proprietary dataset of raw electrocardiographic signals obtained from 164 subjects. The cohort consists of two classes:

1. *Healthy Control Group (Label 0):* 72 subjects with no history of cardiovascular pathology.

2. *High-Risk Group (Label 1):* 92 subjects with cardiovascular disorders or exhibiting high-risk factors.

The class distribution is visualized on Fig. 2. The dataset exhibits a mild class imbalance (44% healthy and 56% high-risk), which reflects real-world clinical scenarios. The signals were recorded at a sampling frequency of 512 Hz, ensuring sufficient temporal resolution for precise morphological analysis.

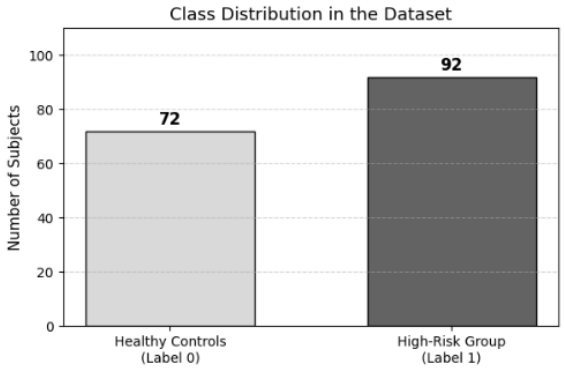

Fig. 2. Distribution of classes in the experimental dataset

Proposed Classification System

Building upon the defined methodological framework, the technical implementation of the proposed solution is divided into two distinct phases. **Phase 1** includes signal processing and feature engineering, initiating with raw ECG data and concluding with the creation of an aggregated patient-level dataset. **Phase 2** focuses on model design and classification, including data splitting, feature selection, hyperparameter tuning and the final model training.

Phase 1: Signal Preprocessing and Feature Engineering. To maximize information extraction while minimizing noise, we developed a multi-stage pipeline (Fig. 3)**.** The workflow converts raw recordings into a structured dataset through the following physiologically grounded steps.

Signal Preprocessing. The input raw ECG data were first processed to remove artifacts such as baseline wander and muscle noise. We utilized a **4th-order Butterworth bandpass filter** designed via the *SciPy signal processing library*. The cutoff frequencies were set to 0.5 Hz and 40 Hz; the lower bound effectively eliminates baseline drift, while the upper bound suppresses high frequency electromyographic noise. To ensure numerical stability, the filter was implemented using Second-Order Sections (SOS). Crucially, we applied a bidirectional filter (*sosfiltfilt* from SciPy) to ensure zero-phase distortion. This preserves the precise shape and timing of the ECG waves, which is essential for accurate morphological analysis.

ISSN 2522-9818 (print)
ISSN 2524-2296 (online)
Сучасний стан наукових досліджень та технологій в промисловості. 2026. № 1 (35)
Innovative technologies and scientific solutions for industries. 2026. No. 1 (35)

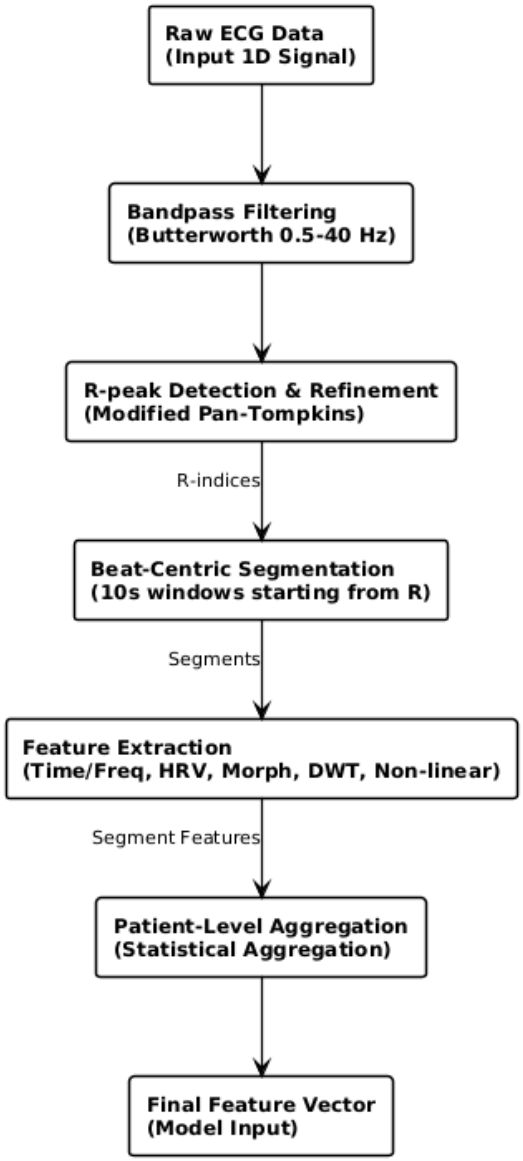

Fig. 3. Workflow of the signal processing and feature engineering pipeline (Phase 1)

R-Peak Detection. Following filtration, precise localization of R-peaks was performed using a multi-stage method inspired by the **Pan-Tompkins algorithm.** The process involved three distinct steps:

1. **Signal Enhancement:** The signal underwent a series of transformations to highlight the QRS complex. This included narrow-band filtration (5-15 Hz), differentiation to highlight slope changes, squaring (y^2) to emphasize high-energy segments, and moving window integration (150 ms window) to smooth the waveform.

2. **Candidate detection:** Peaks were identified on the transformed signal using an adaptive thresholding technique. The threshold was dynamically set as the signal median plus four times the **Median Absolute Deviation (MAD)**. This approach ensures robustness against outliers and varying signal amplitudes. A minimum distance of 250 ms was enforced to prevent multiple detections within a single heartbeat.

3. **Refinement and Relocalization:** Since the enhancement steps can slightly shift peak positions, each candidate was relocalized in the original filtered signal. The algorithm searched for the precise local maximum within a ±100 ms window around the candidate. Finally, a refractory period check eliminated any duplicate detections that were physiologically impossible (closer than 250 ms).

Beat-centric Segmentation and Feature Extraction. Leveraging the validated R-peak indices, the continuous ECG signal was segmented into fixed 10-second windows. Crucially, we employed a beat-centric alignment strategy, ensuring that each segment initiates exactly at an R-peak. This synchronization guarantees temporal consistency across all samples. Subsequently, a comprehensive feature set was extracted from each segment, spanning multiple physiological domains:

- **Time-Domain Statistics:** Basic statistical descriptors of the signal amplitude, including mean, variance, skewness, kurtosis and signal energy.
- **Frequency-Domain Metrics:** Computed using Welch's method to estimate Power Spectral Density (PSD). Key features include Spectral Entropy, Dominant Frequency and relative power in Low-Frequency (LF) and High-Frequency (HF) bands.
- **Heart Rate Variability (HRV):** Derived from the inter-beat (R-R) intervals within the segment. We calculated standard metrics such as SDNN (total variability), RMSSD (short-term variability) and pNN50 to capture autonomic nervous system activity.
- **Morphological Features:** Direct measurements of the cardiac cycle shape, specifically the mean R-peak amplitude and the average QRS complex width.
- **Wavelet Features:** To analyze non-stationary characteristics, we utilized the Discrete Wavelet Transform (DWT). The signal was decomposed using the Daubechies 4 (db4) wavelet, and energy statistics were computed for different decomposition levels.
- **Non-linear Features:** We calculated Sample Entropy to quantify the complexity and regularity of the time series data.

Patient-Level Aggregation: The final step involved transitioning from segment-level data to a patient-level profile. The extracted segments were grouped by patient

Сучасний стан наукових досліджень та технологій в промисловості. 2026. № 1 (35) *ISSN 2522-9818 (print)*
Innovative technologies and scientific solutions for industries. 2026. No. 1 (35) *ISSN 2524-2296 (online)*

ID. To create a stable feature vector that minimizes the impact of local artifacts, we applied statistical aggregation. For every feature column F_j (e.g., RMSSD), four descriptors were computed across all segments of the patient: Mean, Standard Deviation, Minimum and Maximum. This resulted in a high-dimensional, robust representation of each patient, ready for the classification phase.

Phase 2: Model Design and Classification. The second phase focused on optimizing the predictive model using the aggregated dataset prepared in Phase 1. The systematic workflow of this phase, covering feature selection and hyperparameter tuning, and validation, is illustrated in **Fig. 4.**

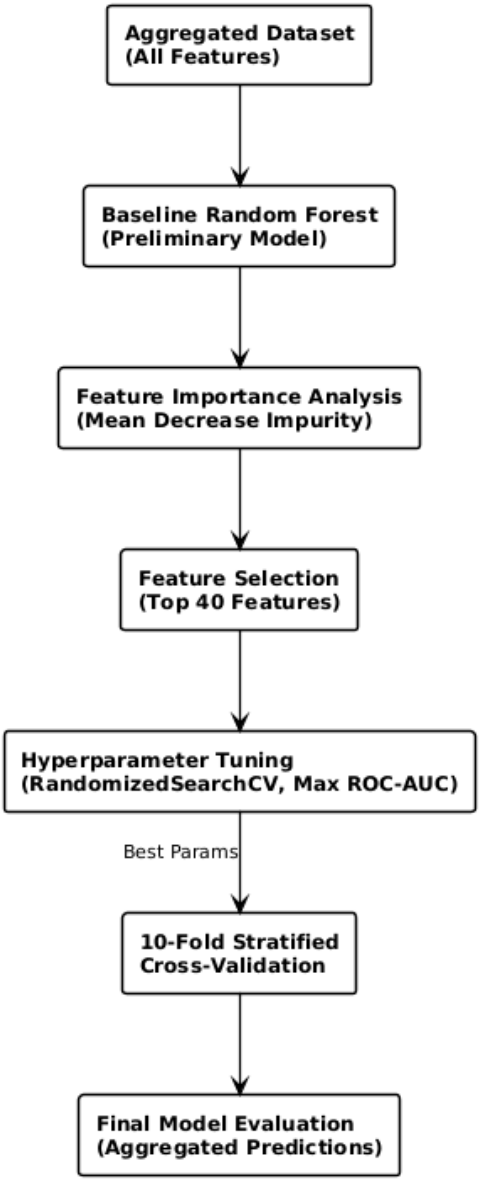

Fig. 4. Workflow of the model optimization and evaluation process (Phase 2)

Feature Selection Strategy: To reduce dimensionality and remove irrelevant predictors, we first trained a baseline Random Forest model on the complete feature set. We utilized the model's built-in feature importance metrics (Mean Decrease Impurity) to quantify the contribution of each feature to the decision-making process. Based on this analysis, we identified and selected the Top 40 most informative features. This step was crucial for focusing the model on the most predictive physiological markers while eliminating noise.

Hyperparameter Tuning: To find the optimal model configuration, we utilized RandomizedSearchCV integrated with 10-fold Stratified Cross-Validation. Unlike exhaustive search methods (e.g., Grid Search), this approach samples a fixed number of parameter settings from specified distributions. This offers a computationally efficient way to explore a high-dimensional hyperparameter space. The optimization process focused on maximizing the ROC-AUC metrics. The search space covered a wide range of hyperparameters, including:

- Number of estimators: 100 – 500
- Maximum tree depth: 2 – 20
- Minimum samples split: 2 - 10
- Minimum samples per leaf: 2 – 10
- Max features: Square root or Log2

Final Evaluation Strategy: To ensure an unbiased assessment of the model's generalization capability, we employed 10-fold Stratified Cross-Validation using optimized hyperparameters. This technique is essential for small and imbalanced datasets, as it preserves the class proportions across all training and validation folds. Final performance metrics were derived by aggregating predictions from the validation sets. Unlike a single train-test split, this approach ensures that every data point is evaluated exactly once as unseen data, providing a realistic estimate of performance while preventing data leakage. Synthetic oversampling (SMOTE) was not applied during this evaluation phase; therefore, the reported metrics (ROC-AUC, Sensitivity, Specificity) accurately reflect the model's performance on the original, real-world class distribution.

Research Results

Comparison of Classification Strategies

To validate the hypothesis that precise beat-centric alignment combined with patient-level aggregation is essential for robust classification, we compared the

ISSN 2522-9818 (print)
ISSN 2524-2296 (online)
Сучасний стан наукових досліджень та технологій в промисловості. 2026. № 1 (35)
Innovative technologies and scientific solutions for industries. 2026. No. 1 (35)

proposed method against a baseline segment-based approach. The baseline model served as a control experiment to quantify the impact of our proposed feature engineering pipeline. In this baseline setup, the processing differed in three key aspects:

1. **Arbitrary Segmentation:** The ECG signals were divided into fixed 10-second windows without synchronization to R-peaks. Consequently, segments could start in the middle of a QRS complex, introducing noise and artifacts.

2. **No Aggregation:** Each segment was classified independently as a separate sample, ignoring the fact that multiple segments belong to the same patient.

3. **Raw Feature Set:** The classification was performed on the full set of extracted features without the specialized selection step used in the final model.

As shown in **Table 1**, this lack of synchronization and patient-level context resulted in suboptimal performance.

Table 1. *Performance comparison between the baseline and the proposed method*

Approach	Recall (0)	Recall (1)	Accuracy	ROC-AUC
Baseline	0.29	0.80	0.57	0.64
Proposed	0.72	0.86	0.80	0.84

The baseline segment-based method demonstrated poor generalization capability, achieving an overall Accuracy of only 57% and an ROC-AUC of 0.64. Crucially, the Recall for the healthy control group was only 0.29, indicating a high rate of false positives caused by local signal artifacts and misalignment. In contrast, the proposed patient-level aggregation strategy significantly improved the model's discriminative power. By statistically aggregating features across the patient's recording, the ROC-AUC increased to 0.84 and Accuracy reached 80%. This improvement in the AUC score confirms that the proposed aggregation method effectively filters out local noise and provides stable subject profile.

Performance of the Final Model

The final Random Forest model was evaluated using Stratified 10-Fold-Cross-Validation. The quantitative results demonstrate the model's robust discriminative capability, achieving an overall Accuracy of 80% and an aggregated ROC-AUC score of 0.84. Detailed performance metrics for each class are summarized in Table 2. The model exhibits a consistent precision profile, achieving a Precision of 0.80 for both Healthy and High-Risk classes. This indicates that when the model predicts a specific label, it is correct in 80% of cases, suggesting a high degree of trustworthiness in positive detections. It should be noted that the support value of 164 in the final row of Table 2 represents the total number of subjects in the dataset.

Table 2. *Detailed classification report of the final optimized mode*

Class	Precision	Recall	F1-Score	Support
Healthy (0)	0.80	0.72	0.76	72
High-Risk (1)	0.80	0.86	0.83	92
Weighted Avg	0.80	0.80	0.80	164

Sensitivity and Specificity Analysis

A critical finding is observed in the Recall metrics. The model achieved a Recall of 0.86 for the High-Risk class, which corresponds to Sensitivity. In medical screening scenarios, maximizing Sensitivity is often prioritized over Specificity to minimize Type II errors (False Negatives). The obtained result confirms that the system successfully identified 86% of all high-risk patients, missing only a small fraction. Conversely the Recall for Healthy class, representing Specificity, reached 0.72. While lower than Sensitivity, this is an acceptable trade-off, as a moderate rate of false alarms (Type I errors) is preferable to missing a pathological condition.

Confusion Matrix Interpretation

The distribution of correct and incorrect predictions is further visualized in the Confusion Matrix (Fig. 5).

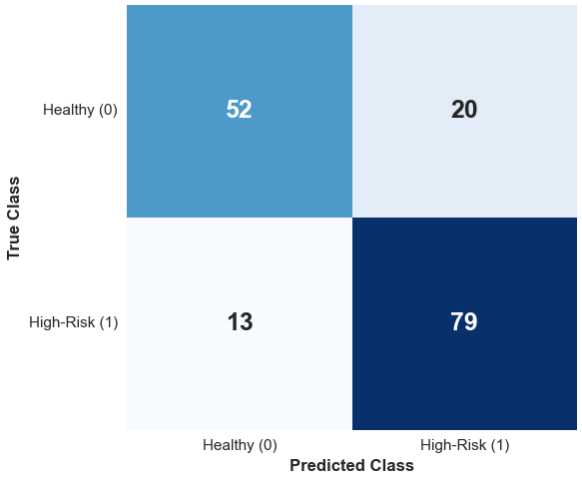

Fig. 5. Confusion Matrix of the final classification model

Сучасний стан наукових досліджень та технологій в промисловості. 2026. № 1 (35) *ISSN 2522-9818 (print)*
Innovative technologies and scientific solutions for industries. 2026. No. 1 (35) *ISSN 2524-2296 (online)*

- True Positives (TP): Out of 92 high-risk subjects, the model correctly identified 79 cases.
- False Negatives (FN): Only 13 high-risk patients were misclassified as healthy.
- True Negatives (TN): The model correctly recognized 52 healthy subjects.
- False Positives (FP): There were 20 cases where healthy individuals were flagged as high-risk.

Feature Importance Analysis

To gain insight into the model's decision-making process, we analyzed feature contributions using the Mean Decrease Impurity metric. The analysis was performed on the final optimized model. As illustrated in Fig. 6, the ranking of the top 10 features reveals a distinct dominance of **Heart Rate Variability (HRV)** metrics. The top five positions were occupied exclusively by aggregated statistics derived from R-R intervals. This confirms that the variability of the heart rate, rather than just the raw morphology, is the strongest predictor of cardiovascular risk in this dataset. Beyond HRV, the model also leveraged **Wavelet-based** features, which capture signal energy in specific frequency bands, and time-domain amplitude descriptors. This hierarchy supports the clinical validity of the proposed approach. The fact that the model prioritized interpretable physiological markers over complex, abstract patterns suggests that the classifier relies on medically relevant phenomena rather than learning noise.

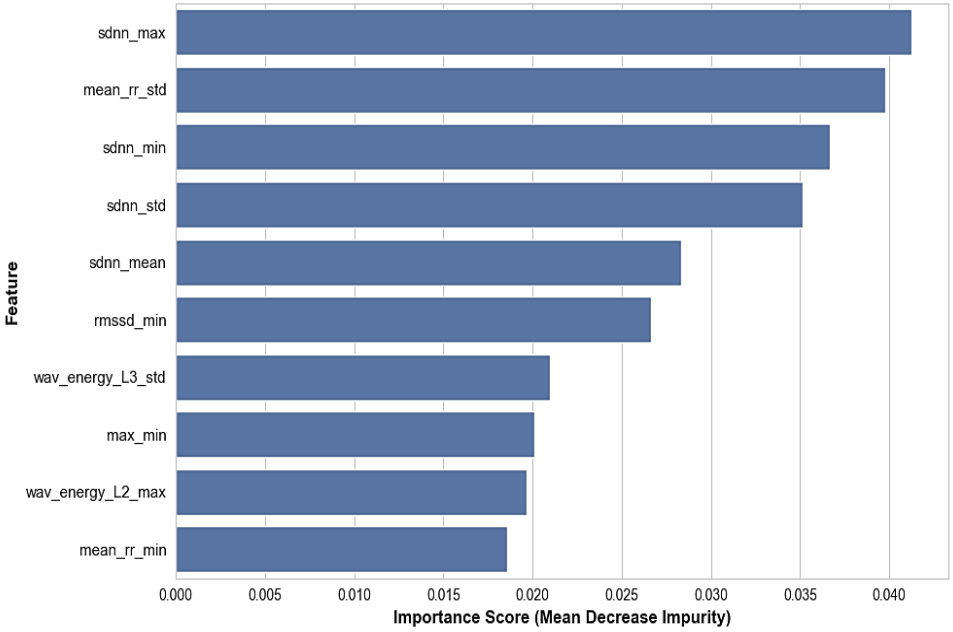

Fig. 6. Top 10 most significant features identified by the Random Forest model

Conclusions

The following results were obtained in this paper:

1. A multi-stage processing pipeline was implemented, utilizing a 4th-order Butterworth filter and a refined Pan-Tompkins algorithm to ensure precise R-peak detection and beat-centric segmentation.

2. A patient-level feature aggregation strategy was developed, which successfully mitigated local signal variability. This approach increased classification accuracy from 57% (baseline segment-based method) to 80%.

3. The Random Forest classifier was optimized using feature selection and hyperparameter tuning. The final model achieved a ROC-AUC of 0.84 and a Sensitivity of 0.86 for high-risk subjects, demonstrating strong screening potential.

4. Feature importance analysis confirmed that Heart Rate Variability (HRV) metrics (SDNN, RMSSD) are the most critical predictors for differentiating between healthy and high-risk subjects, confirming the clinical interpretability of the proposed model.

Despite the encouraging results, this study has several limitations that should be acknowledged. First, the experimental evaluation was conducted on a relatively small, single-center proprietary dataset, which may limit the generalizability of the findings to broader

ISSN 2522-9818 (print) Сучасний стан наукових досліджень та технологій в промисловості. 2026. № 1 (35)
ISSN 2524-2296 (online) Innovative technologies and scientific solutions for industries. 2026. No. 1 (35)

and more heterogeneous populations. Second, although strict patient-level cross-validation was applied to prevent data leakage, external validation on an independent cohort is required to fully assess the robustness of the proposed framework. Third, the current model relies on tree-ensemble feature importance, which, while informative, provides only a global view of decision drivers.

Future work will therefore focus on two complementary directions. From a modeling perspective, the proposed patient-level feature representation creates a natural foundation for fuzzy decision trees and fuzzy classifiers, which can translate aggregated physiological features into explicit, linguistically interpretable rules, enhancing clinical transparency and trust [18]. From an explainability perspective, integrating local explanation techniques such as SHAP would enable subject-specific analysis of risk factors, supporting individualized clinical decision-making. Together, these extensions aim to combine robust patient-level aggregation, formal uncertainty handling, and fine-grained interpretability, further advancing ECG-based risk prediction in small-data clinical settings.

Conflict of interest

The authors declare that they have no conflicts of interest, including financial, personal, authorial, or any other nature, that could influence the research or the results published in this article.

Funding

The study was conducted without financial support.

Data availability

The manuscript has no associated data.

Use of artificial intelligence

The authors confirm that they did not use artificial intelligence technology in the creation of this paper.

References

1. "World Heart Report 2023: Confronting the World's Number One Killer. Geneva, Switzerland. World Heart Federation". available at: https://medbox.org/document/world-heart-report-2023-confronting-the-worlds-number-one-killer
2. Global, Regional, and National Burden of Cardiovascular Diseases and Risk Factors in 204 Countries and Territories, 1990-2023. *JACC Journals*, Vol. 86, No. 22, 2025, pp. 2167-2243. DOI: https://doi.org/10.1016/j.jacc.2025.08.015
3. Berkaya, S. K. et al. (2018), "A Survey on ECG Analysis", *Biomedical Signal Processing and Control*, vol. 43, pp. 216-235, doi: https://doi.org/10.1016/j.bspc.2018.03.003
4. Yong O., Yang L., Kardos A., Zhao Y. (2026), "Non-invasive cardiovascular and vital signs monitoring techniques: review, challenges, and perspectives, Measurement", Vol. 258, part E, 119472 p. DOI: https://doi.org/10.1016/j.measurement.2025.119472
5. Ghasad, P.P., Vegivada, J.V.S,. Kamble, V.M., et al. (2025), "A systematic review of automated prediction of sudden cardiac death using ECG signals", *Physiological Measurement*, Vol. 46. DOI: https://doi.org/10.1088/1361-6579/ad9ce5
6. Wu, Z., Guo, C. (2025), "Deep learning and electrocardiography: systematic review of current techniques in cardiovascular disease diagnosis and management", *BioMed Eng OnLine*, Vol. 24, No. 23, DOI: https://doi.org/10.1186/s12938-025-01349-w
7. Singh, A.K, Krishnan, S. (2023), "ECG signal feature extraction trends in methods and applications", *Biomed Eng Online*, Vol. 22(1):22. DOI: https://doi.org/10.1186/s12938-023-01075-1
8. Kutlu, Y., Kuntalp, D. (2012), "Feature Extraction for ECG Heartbeats Using Higher Order Statistics of WPD Coefficients", *Computer Methods and Programs in Biomedicine*, Vol. 105, No. 3, pp. 257-267, DOI: https://doi.org/10.1016/j.cmpb.2011.10.002
9. Pan, J., Tompkins, W. J. (1985), "A Real-Time QRS Detection Algorithm", *IEEE Trans. on Biomedical Engineering*, Vol. 32, No. 3, pp. 230-236, DOI: https://doi.org/10.1109/TBME.1985.325532
10. Zhai, D., Bao, X., Long, X., Ru, T. and Zhou, G. (2023), "Precise Detection and Localization of R-Peaks From ECG Signals", *Mathematical Biosciences and Engineering*, Vol. 20, No. 11, pp. 19191-19208, DOI: https://doi.org/10.3934/mbe.2023848
11. Safdar, M.F., Nowak, R.M., Pałka, P. (2023), "Pre-Processing techniques and artificial intelligence algorithms for electrocardiogram (ECG) signals analysis: A comprehensive review", *Computers in Biology and Medicine*, Vol. 170, 107908 p. DOI: https://doi.org/10.1016/j.compbiomed.2023.107908
12. Breiman, L. (2001), "Random Forests", *Machine Learning*, Vol. 45, No. 1, pp. 5-32, DOI: https://doi.org/10.1023/A:1010933404324
13. Rabcan, J., Zaitseva, E., Levashenko, V., Kvassay, M. (2025), "Advancing ECG Signal Classification with a Fuzzy Classifier Approach", *IEEE Access*, Vol.13, 2025, pp. 83840-83856. DOI: https://doi.org/10.1109/ACCESS.2025.3568086

Сучасний стан наукових досліджень та технологій в промисловості. 2026. № 1 (35) *ISSN 2522-9818 (print)*
Innovative technologies and scientific solutions for industries. 2026. No. 1 (35) *ISSN 2524-2296 (online)*

14. Zaitseva, E., Rabcan, J., Levashenko, V., Kvassay, M. (2023), "Importance analysis of decision-making factors based on fuzzy decision trees", *Applied Soft Computing*, Vol. 134, 109988 p. DOI: https://doi.org/10.1016/j.asoc.2023.109988
15. Levashenko, V., Zaitseva, E. (2002), "Usage of New Information Estimations for Induction of Fuzzy Decision Trees". In: Yin, H., Allinson, N., Freeman, R., Keane, J., Hubbard, S. (eds) Intelligent Data Engineering and Automated Learning - IDEAL 2002. IDEAL 2002. *Lecture Notes in Computer Science*, Vol 2412. Springer, Berlin, Heidelberg, DOI: https://doi.org/10.1007/3-540-45675-9_74
16. Zaitseva, E., Levashenko, V., Puuronen, S. (2007), "Fuzzy classifier based on fuzzy decision tree", *Proc. of the Int. Conf. on Computer as a Tool* (EUROCON), 2007, pp. 823-827, DOI: https://doi.org/10.1109/EURCON.2007.4400614
17. Klabunde, R.E. (2021), "Normal Sinus Rhythm", *Cardiovascular Physiology Concepts*. available at: https://cvphysiology.com/arrhythmias/a009
18. Zaitseva, E., Levashenko, V., Kvassay, M., Deserno T. (2026), "Reliability estimation of healthcare systems using Fuzzy Decision Trees", *Proc. of the Fed. Conf. on Computer Science and Information Systems* (FedCSIS), 2016, pp. 331-340, DOI: https://doi.org/10.15439/2016F150

Received (Надійшла) 01.01.2026
Accepted for publication (Прийнята до друку) 10.02.2026
Publication date (Дата публікації) 30.03.2026

Відомості про авторів / About the Authors

Крайчі Олександр – аспірант, Жилінський університет, кафедра інформатики, факультет управлінських наук та інформатики; Жиліна, Словаччина;
Alexander Krajči– postgraduate student, University of Žilina, Department of Informatics, Faculty of Management Science and Informatics; Žilina, Slovakia;
e-mail: alexander.krajci@st.fri.uniza.sk
ORCID ID: https://orcid.org/0009-0003-7951-3203
Scopus ID: https://www.scopus.com/authid/detail.uri?authorId=60174647200

Людмила Сидоренко – Державний університет медицини та фармації імені Ніколае Тестеміцану, доцент кафедри молекулярної біології та генетики людини; Кишинів, Молдова;
Ludmila Sidorenko– State University of Medicine and Pharmacy “Nicolae Testemițanu”, Associate Professor of the Department of Molecular Biology and Human Genetics; Chișinău, Moldova;
e-mail: ludmila.sidorenco@usmf.md
ORCID ID: https://orcid.org/0000-0003-0382-4542
Scopus ID: https://www.scopus.com/authid/detail.uri?authorId=57190659134

Барковська Олеся Юріївна – кандидат технічних наук, доцент, Харківський національний університет радіоелектроніки, доцент кафедри Електронних обчислювальних машин; Харків, Україна;
Olesia Barkovska – Ph.D (Engineering Sciences), Associate Professor, Kharkiv National University of Radio Electronics, Associate Professor of the Department of Electronic Computers; Kharkiv, Ukraine;
e-mail: olesia.barkovska@nure.ua
ORCID ID: https://orcid.org/0000-0001-7496-4353
Scopus ID: https://www.scopus.com/authid/detail.uri?authorId=24482907700

НАДІЙНЕ ПРОГНОЗУВАННЯ СЕРЦЕВО-СУДИННОГО РИЗИКУ НА НЕВЕЛИКИХ НАБОРАХ ДАНИХ З ВИКОРИСТАННЯМ ПЕРЕДОВОЇ ІНЖЕНЕРІЇ ОЗНАК

Актуальність. Серцево-судинні захворювання залишаються провідною причиною смертності в усьому світі, що створює високий попит на автоматизовані діагностичні системи. Однак розробка надійних моделей машинного навчання для аналізу електрокардіограми (ЕКГ) часто ускладнюється наявністю лише невеликих за масштабом та незбалансованих наборів даних, що обмежує ефективність підходів глибокого навчання. **Об'єктом дослідження** є процес автоматизованої обробки та класифікації електрокардіографічних сигналів для діагностичних цілей. **Предметом дослідження** є методи

ISSN 2522-9818 (print) *Сучасний стан наукових досліджень та технологій в промисловості. 2026. № 1 (35)*
ISSN 2524-2296 (online) *Innovative technologies and scientific solutions for industries. 2026. No. 1 (35)*

вилучення ознак на основі серцево-судинних захворювань, стратегії агрегації на рівні пацієнта та алгоритми машинного навчання для прогнозування серцево-судинного ризику. **Метою цієї статті** є розробка та оцінка надійної системи класифікації, оптимізованої для невеликих наборів даних, яка підвищує точність прогнозування шляхом використання агрегації ознак на рівні пацієнта та моделей машинного навчання, що пояснюються. **Результати дослідження**. У дослідженні пропонується конвеєр, що починається зі стандартної попередньої обробки сигналу, а потім виконується точне виявлення R-піку та сегментація на основі серцево-судинних захворювань. Фізіологічні ознаки (ВСР, вейвлет, морфологічні) потім витягуються з окремих сегментів та статистично агрегуються на рівні пацієнта. Експерименти на наборі даних із 164 суб'єктів показали, що запропонована стратегія агрегації на рівні пацієнта значно перевершує традиційний аналіз на основі сегментів. Остаточна модель випадкового лісу досягла ROC-AUC балу 0,84. Аналіз важливості ознак підтвердив критичну роль показників варіабельності серцевого ритму (HRV), зокрема SDNN та RMSSD, у диференціації здорових суб'єктів та суб'єктів з високим ризиком.

Ключові слова: класифікація; прогнозування ризику; важливість ознак; машинне навчання; ЕКГ; випадковий ліс.

Бібліографічні описи / Bibliographic descriptions

Крайчі О., Сидоренко Л., Барковська О.Ю. Надійне прогнозування серцево-судинного ризику на невеликих наборах даних з використанням передової інженерії ознак. *Сучасний стан наукових досліджень та технологій в промисловості*. 2026. № 1 (35). С. 55–64. DOI: https://doi.org/10.30837/2522-9818.2026.1.055

Krajči, A., Sidorenko, L., Barkovska, O. (2026), "Predicting Risks of Cardiovascular Disease on Small Datasets using Feature Engineering", *Innovative Technologies and Scientific Solutions for Industries*, No. 1 (35), P. 55–64. DOI: https://doi.org/10.30837/2522-9818.2026.1.055