

Explainable Feature-based Machine Learning for Patient-independent ECG Arrhythmia Classification

Leila Sahraoui¹, Abdelaziz Frih², and Hacene Frih^{2,*}

¹ Department of Electronics, Faculty of Technology, Badji Mokhtar Annaba University,
P.O. Box 12, 23000, Annaba, Algeria

² Department of Biology, Environmental Biosurveillance Laboratory, Faculty of Sciences,
Badji Mokhtar University, Annaba, Algeria

Email: leila.sahraoui@univ-annaba.dz (L.S.); zizoufrih5@gmail.com (A.F.); hacene.frih@univ-annaba.dz (H.F.)

Manuscript received July 22, 2026; revised August 21, 2026; accepted September 8, 2026; published October 26, 2026

*Corresponding author

Abstract—Automated Electrocardiogram (ECG) heartbeat classification plays an important role in arrhythmia diagnosis and clinical decision-support systems. Although deep-learning approaches have recently achieved remarkable performance, many require substantial computational resources and lack interpretability. This study proposes an interpretable feature-based machine-learning framework for automated heartbeat classification using the MIT-BIH Arrhythmia Database. Twenty physiologically meaningful morphological and temporal features were extracted from each heartbeat. Random Forest, XGBoost, and CatBoost, were evaluated under a rigorous patient-independent inter-record validation protocol. Random Forest achieved the highest overall accuracy (89.34%), whereas CatBoost obtained the highest weighted precision (90.90%), weighted F1-score (89.28%), and macro F1-score (40.77%). ROC analysis demonstrated strong discrimination for ventricular ectopic beats (AUC = 0.986) and fusion beats (AUC = 0.937), whereas supraventricular ectopic beats remained challenging (AUC = 0.653) because of their morphological similarity to normal beats and severe class imbalance. Feature-importance analysis identified the previous RR interval and waveform statistical descriptors among the most influential predictors. These findings demonstrate that interpretable feature-based machine learning can provide computationally efficient ECG classification under patient-independent evaluation, although external validation and real-time implementation studies are required before clinical deployment.

Index Terms—Electrocardiogram (ECG), arrhythmia classification, machine learning, random forest, patient-independent validation, feature engineering, explainable artificial intelligence, biomedical signal processing

I. INTRODUCTION

Cardiovascular diseases remain the leading cause of mortality worldwide and represent a major public health challenge. Among these disorders, cardiac arrhythmias constitute a heterogeneous group of abnormalities affecting the generation or conduction of electrical impulses within the heart. Depending on their severity,

arrhythmias may lead to syncope, heart failure, stroke, or sudden cardiac death. Consequently, early and accurate detection of abnormal cardiac rhythms is essential for effective clinical management and risk reduction [1].

The Electrocardiogram (ECG) remains the gold-standard non-invasive technique for monitoring cardiac electrical activity and diagnosing arrhythmias. ECG recordings provide valuable information regarding cardiac rhythm, conduction pathways, and waveform morphology, enabling clinicians to identify a wide range of cardiac abnormalities. However, manual interpretation of long-term ECG recordings is time-consuming, labor-intensive, and subject to inter-observer variability, particularly in ambulatory monitoring and large-scale screening programs.

Recent advances in Artificial Intelligence (AI), Machine Learning (ML), and Deep Learning (DL) have stimulated substantial interest in automated ECG analysis. Numerous studies have demonstrated the potential of intelligent algorithms to improve the accuracy and efficiency of arrhythmia detection while reducing clinician workload. Ansari *et al.* [2] reviewed the major deep-learning architectures applied to ECG classification, including Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), transformers, and hybrid attention-based frameworks, highlighting their remarkable performance on benchmark datasets. Katal *et al.* [3] further confirmed the growing potential of CNN-based architectures for automated cardiac rhythm analysis, while Jaya Prakash *et al.* [4] reviewed advances in both machine-learning and deep-learning methods for ECG beat classification, emphasizing continuous improvements in feature engineering, classification accuracy, and computational efficiency. Similarly, Xiao *et al.* [5] reported that the MIT-BIH Arrhythmia Database remains the most frequently used benchmark dataset for automated ECG classification and highlighted the increasing number of studies reporting classification accuracies exceeding 95%. Beyond arrhythmia classification, artificial intelligence has also demonstrated considerable potential

for the diagnosis and risk stratification of a broad range of cardiovascular diseases, further emphasizing the growing clinical relevance of automated ECG analysis [6].

Recent work published in IJEETC has also demonstrated the value of combining machine-learning and deep-learning approaches for cardiovascular diagnosis. Jamage *et al.* [7] evaluated several machine-learning classifiers using structured clinical data and developed a hybrid CNN–RNN architecture for multiclass ECG analysis, highlighting the potential of intelligent models for automated cardiac assessment.

In recent years, several sophisticated architectures have achieved exceptionally high performance. Bai *et al.* [8] proposed a hybrid CNN–BiGRU architecture with multi-head attention and reported an accuracy of 99.41% on the MIT-BIH database. These findings illustrate the remarkable predictive capability of modern deep-learning models for automated arrhythmia classification.

Despite these impressive results, growing concerns have emerged regarding the comparability and clinical relevance of many published studies. Recent systematic reviews have shown that a substantial proportion of ECG classification studies fail to adhere strictly to the recommendations of the Association for the Advancement of Medical Instrumentation (AAMI), particularly regarding heartbeat grouping and evaluation methodology. Chazal *et al.* [9] highlighted that many investigations continue to employ evaluation protocols that do not adequately reflect real-world clinical deployment conditions. These methodological issues have raised increasing concerns regarding the reliability, reproducibility, and fair comparison of published classification results.

One of the most frequently reported methodological limitations concerns the use of random beat-level splitting strategies. Under such protocols, heartbeats extracted from the same patient may appear simultaneously in both training and testing datasets, introducing information leakage and artificially inflating classification performance. Several recent reviews have emphasized that these approaches often produce overly optimistic results and hinder fair comparison between competing algorithms. To address this limitation, patient-independent inter-record validation has been increasingly recommended. Under this paradigm, ECG records used for testing are completely excluded from model training, providing a more realistic estimate of generalization performance on previously unseen patients. Zhou *et al.* [10] demonstrated the importance of inter-patient evaluation when addressing class imbalance problems in arrhythmia classification, while Silva *et al.* [11] identified inter-patient validation as one of the key requirements for clinically relevant ECG classification studies.

Although deep-learning approaches currently dominate the literature, feature-based machine-learning methods remain attractive because of their lower computational requirements, improved interpretability, and ease of deployment in resource-constrained environments.

Recent studies have demonstrated that carefully engineered ECG features remain highly competitive while providing physiologically interpretable information and substantially reducing computational complexity. Kirkbas [12] demonstrated that feature-based ECG classification remains highly effective even when using a limited number of recordings, while Atwa *et al.* [13] emphasized the growing importance of interpretable artificial intelligence approaches in cardiovascular diagnostics.

The selection of morphological and RR-interval descriptors in the present study was guided by both physiological relevance and previous evidence reported in the literature. Morphological features characterize waveform shape, amplitude variation, and QRS-complex properties associated with ventricular depolarization, whereas RR-interval features describe beat-to-beat temporal dynamics and rhythm irregularity. Together, these complementary descriptors provide a compact yet clinically meaningful representation of cardiac electrical activity, enabling accurate and interpretable arrhythmia classification while maintaining low computational complexity.

Another important challenge concerns class imbalance and morphological overlap between heartbeat categories. Minority classes such as supraventricular ectopic beats and fusion beats often remain difficult to classify accurately, even for advanced deep-learning models. Recent reviews continue to identify class imbalance, generalization ability, interpretability, and reproducibility as major unresolved challenges in automated arrhythmia classification.

From a biomedical signal-processing perspective, this study emphasizes the extraction of physiologically meaningful descriptors rather than relying exclusively on end-to-end learning from raw ECG signals. Following heartbeat segmentation and annotation-guided preprocessing, each ECG beat was transformed into a compact set of physiologically meaningful descriptors. These descriptors combine morphological characteristics of the QRS complex with temporal rhythm information derived from RR intervals. This feature-engineering strategy reduces data dimensionality while preserving clinically relevant information associated with ventricular depolarization and cardiac rhythm dynamics.

Consequently, the proposed framework constitutes a compact biomedical signal-processing pipeline that supports interpretable machine-learning analysis while limiting computational complexity.

Although the individual ECG descriptors used in this study are well established, the contribution of the present work lies in their systematic integration within a unified and transparent evaluation framework. Specifically, the proposed approach combines 20 physiologically interpretable morphological and temporal features with strict patient-independent inter-record validation, comparative evaluation of three ensemble machine-learning classifiers, class-specific performance analysis, feature-importance assessment, and feature-space sensitivity analysis. Rather than introducing new

ECG descriptors, this study therefore advances existing feature-based approaches by jointly emphasizing prevention of patient-level information leakage, interpretation of model decisions, explicit evaluation of minority heartbeat classes, and numerical stability of the extracted feature representation. This combination provides a more comprehensive assessment of feature-based ECG classification than evaluations based predominantly on global classification accuracy.

Motivated by these considerations, the present study evaluates three widely used ensemble machine-learning algorithms—Random Forest, XGBoost, and CatBoost—for automated ECG heartbeat classification using the MIT-BIH Arrhythmia Database. Heartbeats were grouped according to the AAMI standard and represented using 20 morphological and temporal descriptors extracted from segmented ECG beats. A strict patient-independent inter-record validation strategy was adopted to prevent information leakage between training and testing records and to assess generalization to previously unseen patients within the MIT-BIH dataset. In addition to overall classification performance, class-specific precision, recall, and F1-score were examined to characterize performance on minority heartbeat categories. Feature-importance analysis was used to identify the most influential ECG descriptors, while a supplementary feature-space sensitivity analysis evaluated classifier stability under controlled perturbations of the extracted features. The study therefore aims to provide a rigorous, transparent, and reproducible assessment of interpretable feature-based machine-learning methods for patient-independent ECG arrhythmia classification.

II. MATERIALS AND METHODS

A. Study Design

This study was designed as a retrospective computational analysis aimed at evaluating the performance of classical machine-learning models for automated ECG arrhythmia classification using the MIT-BIH Arrhythmia Database. The methodological framework was based on expert-annotated ECG recordings, heartbeat segmentation, extraction of morphological and RR-interval features, patient-independent inter-record validation, machine-learning classification, and feature-space sensitivity analysis.

The main objective was to compare the classification performance of Random Forest, XGBoost, and CatBoost models under a validation strategy that prevents information leakage between training and testing records. A secondary objective was to assess the sensitivity of trained classifiers to additive perturbations introduced into the extracted feature space.

The overall methodological workflow adopted in this study is illustrated in Fig. 1.

Workflow of the proposed ECG arrhythmia classification framework, including ECG data acquisition from the MIT-BIH Arrhythmia Database, expert annotation processing, heartbeat segmentation, feature

extraction, patient-independent inter-record validation, machine-learning classification, performance evaluation, and feature-space sensitivity analysis.

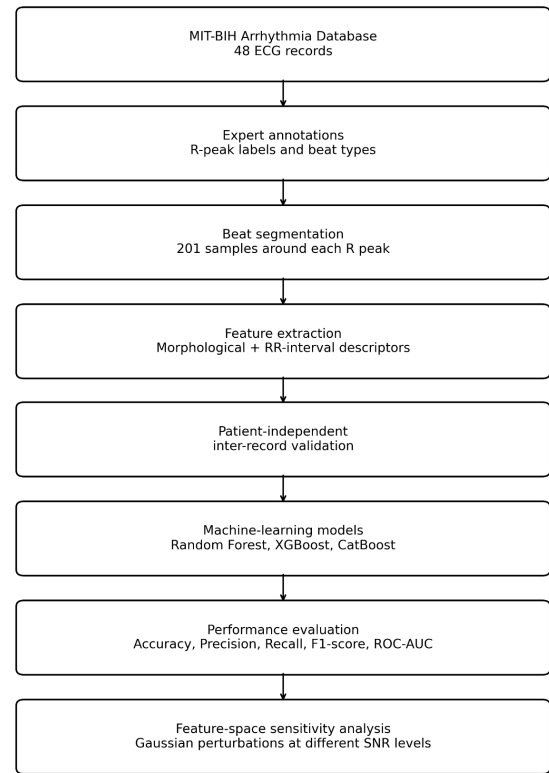


Fig. 1. Workflow of the proposed ECG arrhythmia classification framework.

B. Dataset

The MIT-BIH Arrhythmia Database was used as the source dataset. It contains 48 half-hour ambulatory ECG recordings obtained from 47 subjects. Each record includes two ECG leads sampled at 360 Hz, with expert annotations identifying R-peak locations and beat types.

The ECG signal files were stored in comma-separated value format, while the corresponding annotation files contained the beat sample position and annotation symbol. The MLII lead was used whenever available because of its frequent use in arrhythmia classification studies and its clear representation of QRS morphology. When MLII was unavailable, the second available ECG lead was used.

C. Heartbeat Annotation Processing and AAMI Grouping

Expert beat annotations were used as reference markers for heartbeat extraction. Non-beat annotations and rhythm-change markers were excluded. Valid annotation symbols were mapped according to the AAMI heartbeat classification scheme into five classes:

- N: normal and bundle branch block-related beats
- S: supraventricular ectopic beats
- V: ventricular ectopic beats
- F: fusion beats
- Q: unknown or paced-related beats

After preprocessing, AAMI grouping, and exclusion of incomplete heartbeat segments, 109,462 valid beats were retained for analysis.

D. Heartbeat Segmentation

Each heartbeat was segmented around the annotated R-peak location. A fixed window of 201 samples was extracted for each beat, consisting of 100 samples before and 100 samples after the annotated R peak, plus the R-peak sample itself. Beats located too close to the beginning or end of the ECG recording were excluded when a complete 201-sample segment could not be obtained.

This segmentation procedure ensured that the extracted segment contained the central QRS complex and surrounding waveform information required for morphological feature extraction.

E. Feature Extraction

A total set of morphological and temporal features was extracted from each segmented heartbeat.

The feature extraction stage constitutes the core biomedical signal-processing component of the proposed framework. Instead of processing raw ECG waveforms directly, clinically meaningful descriptors summarizing waveform morphology, statistical characteristics, and cardiac rhythm dynamics were extracted from each heartbeat. This transformation converts high-dimensional ECG signals into a compact feature space that preserves diagnostically relevant information while substantially reducing computational cost, memory requirements, and execution time compared with deep-learning models trained on raw ECG signals.

1) Morphological features

For each 201-sample heartbeat segment, the following global statistical and waveform descriptors were computed:

- Mean amplitude
- Standard deviation
- Minimum amplitude
- Maximum amplitude
- Peak-to-peak amplitude
- Root mean square
- Signal energy
- Skewness
- Kurtosis
- Median
- Median absolute deviation

Additional QRS-related descriptors were also extracted:

- R-peak amplitude
- Q-wave approximate amplitude
- S-wave approximate amplitude
- QRS area
- Approximate QRS width

For QRS-related feature extraction, the annotated R peak was positioned at the center of each 201-sample heartbeat segment (sample 100). The R-peak amplitude was defined as the ECG amplitude at this central sample. Approximate Q- and S-wave amplitudes were determined

as the minimum signal amplitudes within 25-sample windows immediately preceding and following the R peak, respectively. Approximate QRS width was estimated using an amplitude threshold defined as 50% of the R-peak amplitude; the width was calculated as the sample distance between the first and last points of the heartbeat segment exceeding this threshold. The QRS-area descriptor was computed as the sum of the absolute amplitudes across the 201-sample heartbeat segment. These operational definitions were applied consistently to all extracted heartbeats.

These features were selected to characterize global waveform morphology, amplitude variability, QRS-related morphology, and distributional shape of the heartbeat segment.

2) RR-interval features

Temporal rhythm descriptors were computed from consecutive annotated R-peak positions. These included:

- Previous RR interval
- Next RR interval
- Local RR interval
- Heart rate

These features were included because arrhythmia classification depends not only on beat morphology but also on beat-to-beat timing and rhythm irregularity. RR intervals were calculated directly from consecutive annotated R-peak sample positions and retained in sample units. The previous RR interval (RR_{prev}) was defined as the difference between the current and preceding R-peak positions, whereas the next RR interval (RR_{next}) was defined as the difference between the subsequent and current R-peak positions. The local RR interval (RR_{local}) was calculated as the arithmetic mean of RR_{prev} and RR_{next} . Heart rate was subsequently estimated as $(360 \times 60)/RR_{local}$, where 360 Hz corresponds to the sampling frequency of the MIT-BIH recordings. For boundary beats without a preceding or subsequent annotated beat, the corresponding RR interval was assigned a value of zero.

F. Final Analytical Dataset

The final analytical dataset contained 109,462 heartbeats described by 20 extracted ECG features, in addition to record identifiers, lead identifiers, and class labels. The final class distribution was strongly imbalanced, with normal beats representing the dominant category.

The final AAMI class distribution was:

- N: 90,603 beats
- Q: 8041 beats
- V: 7235 beats
- S: 2781 beats
- F: 802 beats

G. Validation Strategy

To evaluate patient-independent generalization within the MIT-BIH Arrhythmia Database, an inter-record validation strategy was applied.

Training and testing datasets were constructed from separate ECG records. No heartbeat extracted from a

testing record was included in the training set. The inter-record split followed the widely adopted patient-independent protocol originally proposed by Chazal *et al.* [9], which has become the reference benchmark for fair comparison among ECG heartbeat classification studies. Training records included 101, 106, 108, 109, 112, 114, 115, 116, 118, 119, 122, 124, 201, 203, 205, 207, 208, 209, 215, 220, 223, and 230. Testing records included 100, 103, 105, 111, 113, 117, 121, 123, 200, 202, 210, 212, 213, 214, 219, 221, 222, 228, 231, and 232.

This separation ensured complete patient independence between model training and evaluation and prevented information leakage between datasets.

This strategy was selected to avoid overly optimistic performance estimates caused by random beat-level splitting, where beats from the same patient or same ECG recording may appear in both training and testing sets.

The patient-independent split resulted in 51,006 training beats and 43,867 testing beats. The detailed allocation of training and testing records is presented in Table I.

TABLE I: PATIENT-INDEPENDENT INTER-RECORD SPLIT USED IN THIS STUDY

Training Records	Testing Records
101	100
106	103
108	105
109	111
112	113
114	117
115	121
116	123
118	200
119	202
122	210
124	212
201	213
203	214
205	219
207	221
208	222
209	228
215	231
220	232
223	
230	

The remaining 14,589 beats originated from records not included in the predefined inter-record protocol and were therefore excluded from model training and evaluation. These records were excluded to strictly preserve the predefined inter-patient evaluation protocol originally proposed by de Chazal *et al.* [9], thereby ensuring fair comparison with previous studies and preventing any potential information leakage.

H. Machine-learning Models

Three supervised machine-learning classifiers were evaluated:

- Random Forest

- XGBoost
- CatBoost

Random Forest was used as a robust ensemble-learning baseline based on multiple decision trees and majority voting. XGBoost and CatBoost were selected as gradient-boosting algorithms capable of modeling nonlinear relationships between morphological and temporal ECG descriptors.

The selection of XGBoost was also motivated by its ability to model complex nonlinear relationships, manage high-dimensional feature sets, and provide feature-importance information. Previous IJEETC research has combined optimized XGBoost classification with feature selection and explainable artificial intelligence to improve predictive performance and model transparency [14].

The models were trained using the extracted feature matrix from the training records and evaluated on independent testing records.

1) Model hyperparameters

Random Forest was trained using 500 decision trees ($n_estimators = 500$) with class balancing enabled ($class_weight = \text{"balanced"}$), a fixed random seed of 42, and parallel computation across available processors.

XGBoost was implemented with 500 boosting iterations ($n_estimators = 500$), maximum tree depth of 6 ($max_depth = 6$), learning rate of 0.05, subsample ratio of 0.8, and column sampling ratio of 0.8. The multiclass objective function ("multi:softmax") was used with a fixed random seed of 42.

CatBoost was trained using 500 boosting iterations ($iterations = 500$), tree depth of 6, learning rate of 0.05, MultiClass loss function, and a random seed of 42.

I. Performance Evaluation

Model performance was assessed using the following metrics:

- Accuracy
- Precision
- Recall
- F1-score
- Receiver operating characteristic area under the curve

Weighted precision, recall, and F1-score were used to account for class imbalance. In addition to weighted metrics, class-specific precision, recall, and F1-score were reported for each AAMI heartbeat category, with particular attention to the minority S and F classes. Multiclass ROC curves were generated using a one-versus-rest approach. Confusion matrices were also computed to identify class-specific misclassification patterns.

J. Feature Importance Analysis

Feature importance was evaluated using the Random Forest classifier. The objective was to determine which morphological and temporal features contributed most strongly to heartbeat classification. Importance values were ranked and visualized to identify the most discriminative ECG descriptors.

K. Feature-space Sensitivity Analysis

A supplementary feature-space sensitivity analysis was performed to evaluate the stability of Random Forest and CatBoost classifiers under additive Gaussian perturbations applied to the extracted feature values.

Perturbations were introduced at different signal-to-noise ratio levels:

- 30 dB
- 20 dB
- 10 dB
- 5 dB
- 0 dB

This analysis was interpreted as sensitivity to perturbations in the feature space, not as a direct simulation of noise contamination in raw ECG signals.

L. Software and Implementation

All analyses were performed using Python. Data processing, feature extraction, model training, evaluation, and figure generation were implemented using Pandas, NumPy, SciPy, Scikit-learn, XGBoost, CatBoost, and Matplotlib.

M. Ethical Considerations

The MIT-BIH Arrhythmia Database is a publicly available anonymized database distributed for research purposes. Therefore, no additional ethical approval or informed consent was required for this retrospective computational study.

III. RESULT AND DISCUSSION

A. Dataset Characteristics

Following heartbeat segmentation, preprocessing, AAMI grouping, and exclusion of incomplete heartbeat segments, a total of 109,462 annotated heartbeats were retained for subsequent analysis. The resulting dataset comprised the five heartbeat categories defined by the Association for the Advancement of Medical Instrumentation (AAMI), namely normal (N), supraventricular ectopic (S), ventricular ectopic (V), fusion (F), and unknown or paced (Q) beats. The distribution of these heartbeat classes is summarized in Table II and illustrated in Fig. 2.

TABLE II: DISTRIBUTION OF AAMI HEARTBEAT CLASSES IN THE MIT-BIH ARRHYTHMIA DATABASE

Class	Number of beats	Percentage (%)
N	90,603	82.77
Q	8041	7.35
V	7235	6.61
S	2781	2.54
F	802	0.73
Total	109,462	100.00

Distribution of the five Association for the Advancement of Medical Instrumentation (AAMI) heartbeat classes used in this study after preprocessing and beat segmentation. A total of 109,462 annotated heartbeats were included. The dataset exhibited

substantial class imbalance, with normal beats (N) accounting for the majority of observations.

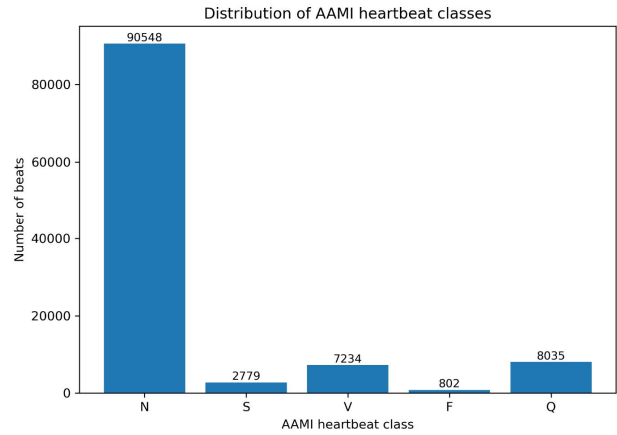


Fig. 2. Distribution of AAMI heartbeat classes in the MIT-BIH arrhythmia database.

Bar chart showing the distribution of heartbeat classes according to the AAMI standard. The dataset is highly imbalanced, with normal beats (N) representing the majority class, while fusion beats (F) and supraventricular ectopic beats (S) constitute minority classes.

As presented in Table II, the dataset exhibited a pronounced class imbalance. Normal beats (N) constituted the dominant category with 90,603 heartbeats (82.77%), whereas ventricular ectopic beats (V) accounted for 7235 (6.61%), unknown beats (Q) for 8041 (7.35%), supraventricular ectopic beats (S) for 2781 (2.54%), and fusion beats (F) for only 802 (0.73%). This distribution reflects the intrinsic characteristics of the MIT-BIH Arrhythmia Database and highlights the limited representation of clinically important abnormal heartbeat classes.

The marked imbalance between majority and minority classes constitutes one of the principal challenges in automated ECG heartbeat classification. Machine-learning algorithms trained on such datasets naturally tend to favor the dominant class during optimization, which may lead to apparently high overall accuracy while simultaneously reducing sensitivity for rare heartbeat categories. This issue is particularly relevant for supraventricular ectopic and fusion beats, whose limited number of training examples restricts the ability of classifiers to learn representative decision boundaries. Similar observations have been highlighted in recent systematic reviews, which identify class imbalance as one of the major obstacles to the development of reliable and clinically applicable ECG classification systems [11, 15].

Beyond class imbalance, the classification task is further complicated by the substantial morphological similarity between several heartbeat categories. In particular, supraventricular ectopic beats often share waveform characteristics with normal sinus beats, making their discrimination considerably more difficult than that of ventricular ectopic beats. Consequently, overall

classification accuracy alone cannot adequately characterize model performance. For this reason, the subsequent analyses include complementary evaluation metrics, including precision, recall, F1-score, Receiver Operating Characteristic (ROC) analysis, and confusion matrices, thereby providing a more comprehensive assessment of classifier behavior across both majority and minority heartbeat classes.

Finally, the combination of severe class imbalance and the patient-independent inter-record validation strategy adopted in this study makes the classification task substantially more demanding than conventional beat-level random partitioning. Since testing records originate from previously unseen patients, classifiers must simultaneously cope with limited training examples for minority classes and inter-patient variability in ECG

morphology and rhythm characteristics. Consequently, the present dataset provides a challenging benchmark for evaluating the performance and patient-independent generalization of machine-learning algorithms within the MIT-BIH Arrhythmia Database.

B. Overall Classification Performance

The comparative performance of the three evaluated machine-learning algorithms is presented in Table III and illustrated in Fig. 3. Overall, the three ensemble-learning models demonstrated comparable performance under the patient-independent inter-record validation protocol, confirming that the extracted morphological and RR-interval descriptors provide a robust representation of heartbeat characteristics despite the challenging evaluation framework.

TABLE III: PERFORMANCE COMPARISON OF MACHINE-LEARNING MODELS UNDER PATIENT-INDEPENDENT INTER-RECORD VALIDATION

Model	Accuracy (%)	Weighted Precision (%)	Weighted Recall (%)	Weighted F1-score (%)	Macro F1-score (%)
Random Forest	89.34	88.27	89.34	88.33	34.88
XGBoost	88.68	90.45	88.68	89.17	40.65
CatBoost	88.85	90.90	88.85	89.28	40.77

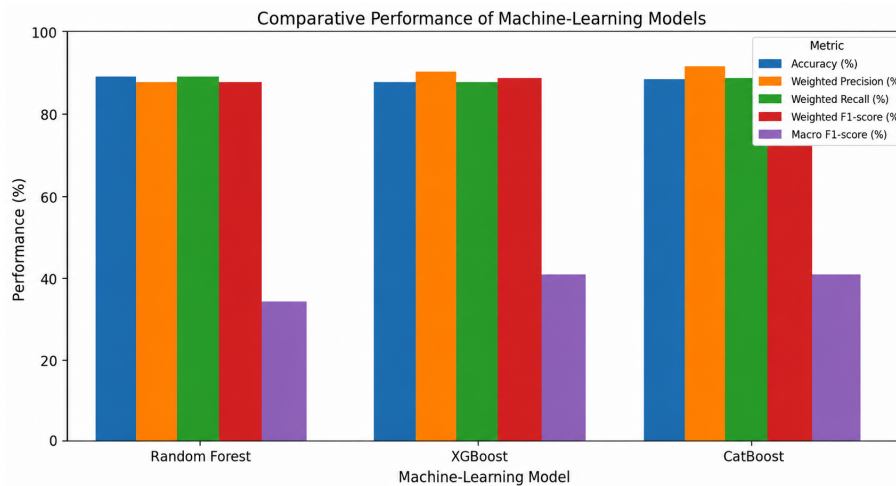


Fig. 3. Comparative performance of Random Forest, XGBoost, and CatBoost under patient-independent inter-record validation in terms of accuracy, weighted precision, weighted recall, weighted F1-score, and macro F1-score.

Classification performance of Random Forest, XGBoost, and CatBoost under patient-independent inter-record validation. Performance was assessed using accuracy, weighted precision, weighted recall, weighted F1-score, and macro F1-score. Random Forest achieved the highest overall accuracy, whereas CatBoost achieved the highest weighted precision, weighted F1-score, and macro F1-score.

Among the evaluated classifiers, Random Forest achieved the highest overall accuracy (89.34%) and weighted recall (89.34%). CatBoost achieved the highest weighted precision (90.90%) and weighted F1-score (89.28%), as well as the highest macro F1-score (40.77%). XGBoost showed comparable performance, with an accuracy of 88.68%, weighted precision of 90.45%, weighted recall of 88.68%, weighted F1-score of 89.17%, and macro F1-score of 40.65%. These results indicate that no single classifier was uniformly superior across all evaluation metrics.

Overall performance remained broadly comparable across the three classifiers, although larger differences were observed for weighted precision and particularly macro F1-score.

Although Random Forest provided the highest overall accuracy, the relatively small performance differences observed among the three classifiers suggest that feature engineering may have contributed more substantially to classification performance than the specific choice of machine-learning algorithm. The extracted descriptors combine complementary information describing both heartbeat morphology and temporal rhythm dynamics, allowing all ensemble-learning models to construct effective decision boundaries despite the pronounced class imbalance and inter-patient variability present in the dataset.

These findings indicate that carefully designed handcrafted ECG features can provide effective arrhythmia classification under patient-independent

evaluation and may serve as a complementary alternative to deep-learning approaches when interpretability and computational efficiency are important considerations.

Unlike end-to-end neural networks that directly process raw ECG waveforms, the present framework relies on physiologically interpretable descriptors with substantially lower computational complexity. Consequently, the proposed methodology offers an attractive compromise between predictive performance, interpretability, and computational efficiency, suggesting potential relevance for embedded systems and single-lead monitoring applications, subject to further external validation.

Direct comparison with recently published studies should nevertheless be interpreted with caution because reported performance strongly depends on the adopted validation strategy. Numerous deep-learning studies have reported classification accuracies exceeding 98–99% on the MIT-BIH Arrhythmia Database. However, many of these investigations employed random beat-level partitioning, allowing heartbeats originating from the same patient to appear in both training and testing datasets. Such protocols introduce information leakage and generally produce overly optimistic estimates of model performance [9, 11]. In contrast, the present study adopted a strict patient-independent inter-record validation protocol in which all testing records were completely excluded from model training. Consequently, the reported performance more realistically reflects the ability of the classifiers to generalize to previously unseen patients, thereby providing a more rigorous estimate of patient-independent performance within the MIT-BIH dataset.

An additional observation concerns the excellent consistency observed across the evaluated classifiers. The

small differences between Random Forest, XGBoost, and CatBoost indicate that classifier performance may be strongly influenced by the quality of the extracted ECG features and the rigor of the validation protocol, rather than by algorithmic complexity alone.

This observation supports the view that transparent feature-engineering strategies can serve as a complementary approach to deep-learning models, particularly when reproducibility, interpretability, and computational efficiency are important considerations [12–14]. However, because no deep-learning model was directly implemented under the same experimental protocol in the present study, no conclusion regarding comparative superiority can be drawn.

Overall, these results indicate that ensemble machine-learning algorithms, when combined with physiologically meaningful ECG descriptors and evaluated under rigorous patient-independent internal validation, can provide useful and interpretable arrhythmia classification performance while maintaining a favorable balance between accuracy and computational efficiency.

C. Class-specific Performance and Minority-class Analysis

Class-specific analysis revealed substantial differences in classification performance across the five AAMI heartbeat categories (Table IV). Normal beats (N) were consistently recognized with high accuracy by all three classifiers, with F1-scores of 94.43% for Random Forest, 93.86% for XGBoost, and 93.95% for CatBoost. This strong performance reflects both the dominant representation of normal beats in the dataset and their relatively consistent morphological and temporal characteristics.

TABLE IV: CLASS-SPECIFIC PERFORMANCE OF THE THREE MACHINE-LEARNING MODELS UNDER PATIENT-INDEPENDENT VALIDATION

Model	Class	Precision (%)	Recall (%)	F1-score (%)	Support
Random Forest	F	10.78	20.42	14.12	377
Random Forest	N	94.70	94.17	94.43	39,316
Random Forest	Q	0.00	0.00	0.00	7
Random Forest	S	11.76	0.34	0.66	1780
Random Forest	V	52.02	87.31	65.20	2387
XGBoost	F	9.96	63.13	17.21	377
XGBoost	N	95.16	92.58	93.86	39,316
XGBoost	Q	0.00	0.00	0.00	7
XGBoost	S	20.18	5.06	8.09	1780
XGBoost	V	78.15	91.08	84.12	2387
CatBoost	F	9.92	66.58	17.27	377
CatBoost	N	95.12	92.82	93.95	39,316
CatBoost	Q	0.00	0.00	0.00	7
CatBoost	S	30.18	4.66	8.08	1,780
CatBoost	V	79.73	89.99	84.55	2387

Note: N: normal beats; S: supraventricular ectopic beats; V: ventricular ectopic beats; F: fusion beats; Q: unknown, paced, or unclassifiable beats according to the adopted AAMI grouping. Precision, recall, and F1-score are reported separately for each heartbeat class.

Ventricular ectopic beats (V) were also effectively identified, particularly by the gradient-boosting models. XGBoost achieved a precision of 78.15%, recall of 91.08%, and F1-score of 84.12%, while CatBoost achieved a precision of 79.73%, recall of 89.99%, and F1-score of 84.55%. Random Forest showed a lower precision of 52.02% but maintained a high recall of 87.31%, resulting in an F1-score of 65.20%. These results

indicate that ventricular ectopic beats exhibit sufficiently distinctive morphological and temporal characteristics to support reliable discrimination under patient-independent evaluation.

In contrast, supraventricular ectopic beats (S) represented the most challenging clinically relevant category. Random Forest correctly identified only 0.34% of S beats, resulting in an F1-score of 0.66%. XGBoost

and CatBoost provided modest improvements, achieving recalls of 5.06% and 4.66% and F1-scores of 8.09% and 8.08%, respectively. These low sensitivities indicate substantial confusion between supraventricular ectopic and other heartbeat categories, particularly normal beats. This difficulty can be attributed to the pronounced class imbalance and to the morphological similarity between supraventricular ectopic and normal beats, for which discrimination frequently depends on subtle temporal rhythm characteristics.

Fusion beats (F) also remained difficult to classify accurately. Although XGBoost and CatBoost achieved relatively high recalls of 63.13% and 66.58%, respectively, their precision remained low at 9.96% and 9.92%, resulting in F1-scores of only 17.21% and 17.27%. Random Forest achieved lower recall (20.42%) and an F1-score of 14.12%. These findings indicate that increased sensitivity to fusion beats was accompanied by a substantial number of false-positive predictions.

The Q category showed zero precision, recall, and F1-score for all three classifiers. However, this result should be interpreted cautiously because only seven Q beats were present in the independent test set. Such an extremely small test support does not permit a reliable assessment of classifier performance for this category. Therefore, the observed zero scores should primarily be regarded as a consequence of the predefined patient-independent dataset split and the resulting insufficient representation of Q beats in the test partition, rather than as evidence of intrinsic algorithmic failure. Future studies should complement fixed patient-independent evaluation with patient-level cross-validation or repeated patient-independent partitioning, where feasible, to obtain more reliable estimates of performance for severely underrepresented heartbeat classes.

Overall, the class-specific results demonstrate that high global accuracy and weighted performance metrics can mask substantial differences in classifier behavior across heartbeat categories. While normal and ventricular ectopic beats were recognized comparatively well, supraventricular ectopic, fusion, and Q beats remained considerably more challenging. These findings emphasize the importance of reporting class-specific metrics alongside global performance measures when evaluating arrhythmia classifiers on strongly imbalanced datasets under patient-independent validation.

D. Receiver Operating Characteristic (ROC) Analysis

To further evaluate class-specific discriminative capability, multiclass receiver operating characteristic (ROC) analysis was performed for the Random Forest classifier, which achieved the highest overall accuracy among the evaluated models. The ROC analysis was conducted using a one-versus-rest strategy in which each heartbeat category was successively considered as the positive class while all remaining categories were grouped as the negative class.

The resulting ROC curves are presented in Fig. 4, while the corresponding Area Under the Curve (AUC) values for each AAMI heartbeat class are summarized in Table V.

TABLE V: CLASS-SPECIFIC AND GLOBAL ROC-AUC VALUES OBTAINED WITH THE RANDOM FOREST CLASSIFIER

ROC-AUC measure	AUC	Test support
F	0.937	377
N	0.828	39,316
Q	0.830	7
S	0.653	1780
V	0.986	2387
Macro-average	0.847	—
Weighted-average	0.830	—

Note: ROC-AUC values were calculated using a one-versus-rest (OvR) strategy on the independent test set. The macro-average assigns equal importance to each AAMI heartbeat class, whereas the weighted-average accounts for class support. The AUC value for the Q class should be interpreted cautiously because only seven Q beats were present in the independent test set.

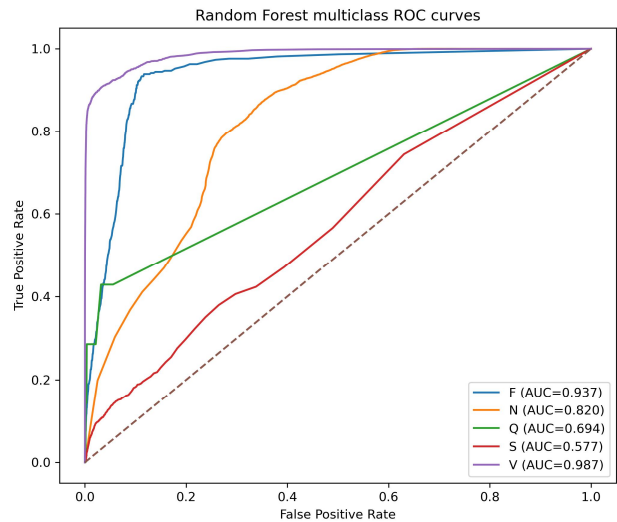


Fig. 4. Class-specific Receiver Operating Characteristic (ROC) curves of the Random Forest classifier on the independent inter-patient test set.

The highest discriminative performance was observed for ventricular ectopic beats (V; AUC = 0.986) and fusion beats (F; AUC = 0.937), whereas supraventricular ectopic beats (S) showed substantially lower discrimination (AUC = 0.653). The AUC obtained for the Q class (0.830) should be interpreted cautiously because only seven Q beats were represented in the test set.

Class-specific ROC analysis revealed substantial differences in discriminative performance across AAMI beat categories. The Random Forest classifier achieved the highest ROC-AUC for class V (0.986), followed by F (0.937), Q (0.830), N (0.828), and S (0.653). These findings indicate particularly strong discrimination of ventricular and fusion beats, whereas supraventricular ectopic beats remained considerably more difficult to distinguish from the remaining classes. Although the Q class yielded an AUC of 0.830, this estimate should be interpreted cautiously because only seven Q beats were available in the independent test set.

The high ROC-AUC observed for class F should be interpreted together with its relatively low precision and F1-score, as ROC-AUC reflects ranking ability across classification thresholds rather than performance at the final multiclass decision rule.

At the global multiclass level, the one-versus-rest ROC analysis yielded a macro-average AUC of 0.847 and a

support-weighted AUC of 0.830. The macro-average result indicates good overall discriminative ability when all AAMI classes are given equal importance. The slightly lower weighted AUC reflects the strong influence of the dominant N class, which represented the majority of the independent test set. These global AUC values indicate that the classifier retained useful overall class-separation ability despite the substantial class imbalance and the strict patient-independent evaluation protocol.

The excellent discrimination obtained for ventricular ectopic beats is consistent with their distinctive ECG morphology. Ventricular beats typically present widened QRS complexes, abnormal depolarization patterns, and altered rhythm dynamics that differ markedly from those of normal sinus beats. These pronounced morphological and temporal differences are effectively captured by the extracted descriptors, particularly the QRS-related features and RR-interval variables, allowing the classifier to establish well-defined decision boundaries.

In contrast, the considerably lower AUC obtained for supraventricular ectopic beats reflects one of the most persistent challenges in automated ECG heartbeat classification. Supraventricular ectopic beats often preserve a QRS morphology that closely resembles that of normal sinus beats because ventricular depolarization usually follows the normal conduction pathway. Consequently, the distinction between these two classes relies primarily on subtle temporal rhythm variations rather than on large morphological differences. This intrinsic similarity substantially reduces class separability and increases the probability of misclassification, particularly under patient-independent evaluation protocols.

The ROC analysis therefore confirms that the proposed feature set provides excellent discrimination for heartbeat categories characterized by pronounced morphological alterations, whereas classes exhibiting substantial waveform overlap remain considerably more challenging. Similar observations have been reported in recent inter-patient ECG classification studies, where ventricular arrhythmias consistently achieve the highest AUC values, while supraventricular ectopic beats remain the weakest-performing class because of both morphological similarity and severe class imbalance [10, 11, 16, 17].

Overall, the multiclass ROC analysis complements the global performance metrics presented previously by demonstrating that classifier performance is not uniformly distributed across heartbeat categories. While the proposed framework achieves excellent discrimination for ventricular arrhythmias, additional methodological developments are still required to improve the recognition of supraventricular ectopic beats, which continue to represent one of the principal limitations of current ECG arrhythmia classification systems.

E. Feature Importance Analysis

To better understand the factors driving heartbeat classification, feature-importance analysis was performed using the Random Forest classifier. The resulting ranking

of the most influential ECG descriptors is presented in Fig. 5 and summarized in Table VI.

TABLE VI: TOP-RANKED ECG FEATURES ACCORDING TO RANDOM FOREST FEATURE-IMPORTANCE ANALYSIS

Rank	Feature	Interpretation
1	RR_prev	Previous RR interval
2	Kurtosis	Global waveform morphology
3	Max amplitude	R/QRS amplitude-related descriptor
4	Skewness	Waveform asymmetry
5	Standard deviation	Signal variability
6	Heart_rate	Temporal rhythm descriptor
7	RR_next	Next RR interval
8	RR_local	Local rhythm variability

Top-ranked ECG features identified by the Random Forest classifier according to feature-importance analysis. Temporal rhythm descriptors, particularly RR_prev, together with higher-order statistical descriptors such as kurtosis and skewness, contributed most strongly to heartbeat classification. The results indicate that both rhythm-related and morphological ECG characteristics play important roles in arrhythmia discrimination.

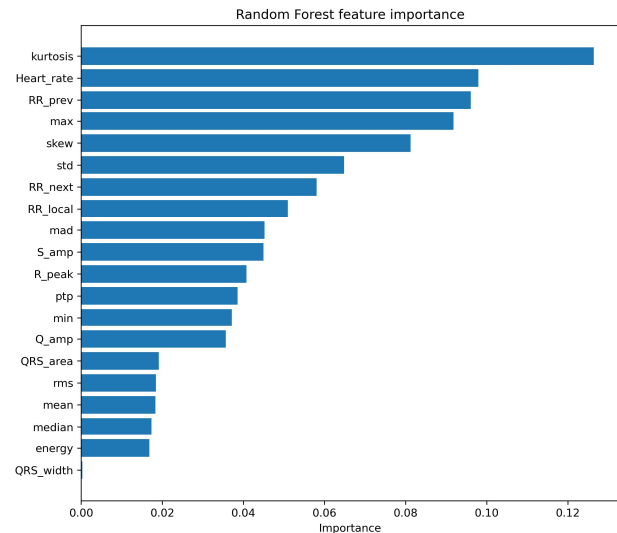


Fig. 5. Feature importance ranking obtained from the random forest model.

Relative importance of the extracted ECG features according to the Random Forest classifier. Temporal rhythm descriptors, particularly RR_prev, together with higher-order statistical descriptors such as kurtosis and skewness, contributed most strongly to heartbeat classification performance.

Among all extracted variables, the previous RR interval (RR_prev) emerged as the most influential predictor of heartbeat class. This finding highlights the major contribution of temporal rhythm information to arrhythmia detection. Since many cardiac arrhythmias are characterized by irregular beat-to-beat timing rather than by pronounced morphological alterations alone, the interval preceding each heartbeat provides valuable physiological information for distinguishing normal and abnormal cardiac rhythms.

The second most important feature was kurtosis, followed by maximum amplitude, skewness, and standard deviation. These descriptors characterize complementary aspects of ECG waveform morphology, including signal distribution, waveform asymmetry, amplitude variation, and overall signal variability. Their high ranking indicates that statistical descriptors of the heartbeat waveform successfully capture morphological changes associated with different arrhythmia types while remaining computationally inexpensive to calculate.

Interestingly, several rhythm-related variables, including *RR_prev*, *RR_next*, *RR_local*, and *Heart rate*, were all ranked among the most informative predictors. This observation demonstrates that temporal rhythm dynamics play a central role in automated heartbeat classification and complement conventional morphological analysis. Rather than relying exclusively on waveform shape, the proposed feature set integrates both electrical activation patterns and beat-to-beat timing information, thereby providing a more comprehensive representation of cardiac electrical activity.

Conversely, QRS-specific descriptors such as QRS width and QRS area contributed less to the final decision process than expected. Although these variables remain clinically relevant for identifying ventricular conduction abnormalities, their lower relative importance suggests that global waveform statistics and rhythm descriptors provide greater discriminatory power within the present patient-independent classification framework. This finding further illustrates that effective heartbeat classification results from the complementary interaction of multiple physiological descriptors rather than from a single dominant ECG characteristic.

The feature-importance analysis also provides valuable insight into the behavior of the Random Forest classifier. Unlike many deep-learning models that operate as “black boxes,” the proposed feature-based framework allows the contribution of each physiological descriptor to be quantified directly. Such interpretability is particularly important in clinical decision-support systems, where understanding the factors influencing automated predictions may increase clinician confidence and facilitate model validation before routine clinical implementation. Recent studies have similarly emphasized the importance of explainable and interpretable artificial intelligence for improving transparency, reliability, and clinical acceptance of automated ECG analysis [13–15, 18].

Overall, these findings demonstrate that the combination of temporal rhythm descriptors and statistical morphological features provides an effective and physiologically meaningful representation of ECG signals. The predominance of RR-based variables confirms that heartbeat timing remains a fundamental determinant of arrhythmia recognition, while higher-order statistical descriptors contribute complementary information describing waveform morphology. Together, these features enable accurate heartbeat discrimination while maintaining a compact, interpretable, and

computationally efficient feature space suitable for practical biomedical engineering applications.

F. Confusion Matrix Analysis

To provide a more detailed assessment of classifier behavior beyond global performance metrics, the confusion matrix obtained with the Random Forest classifier is presented in Fig. 6. Unlike overall accuracy or ROC analysis, the confusion matrix allows visualization of class-specific prediction outcomes and identifies the principal sources of misclassification among the five AAMI heartbeat categories.

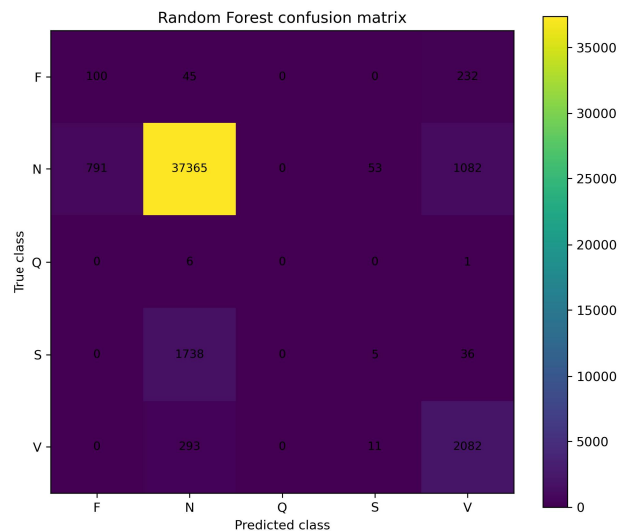


Fig. 6. Confusion matrix of the random forest classifier under patient-independent inter-record validation.

Confusion matrix illustrating the classification outcomes of the Random Forest model across the five AAMI heartbeat classes. Most classification errors involved supraventricular ectopic beats (S), which were frequently misclassified as normal beats (N), reflecting the morphological similarity between these categories.

The classifier demonstrated excellent recognition of the dominant normal heartbeat (N) class, with the vast majority of normal beats correctly assigned to their true category. This result is expected because normal beats represent more than 82% of the entire dataset and therefore provide abundant examples for model training. In addition, normal sinus beats exhibit relatively consistent morphological and temporal characteristics, facilitating reliable discrimination from most abnormal heartbeat classes.

Similarly, ventricular ectopic beats (V) were classified with high accuracy. Their characteristic ECG morphology, including abnormal ventricular depolarization, widened QRS complexes, and distinctive rhythm alterations, provides strong discriminatory information for the extracted morphological and temporal descriptors. These observations are fully consistent with the high ROC-AUC value obtained for this class and confirm that ventricular arrhythmias are among the easiest heartbeat categories to identify automatically.

In contrast, the most pronounced classification errors involved supraventricular ectopic beats (S). A considerable proportion of S beats were incorrectly classified as normal beats, indicating that the extracted features were not always sufficient to separate these two categories under patient-independent evaluation. This behavior is not unexpected because supraventricular ectopic beats frequently preserve a narrow QRS complex and ventricular activation pathway similar to those of normal sinus beats. Consequently, the distinction between these classes often depends on subtle rhythm variations that may differ substantially between patients.

Although fusion beats (F) were represented by only a small number of observations, the classifier still achieved satisfactory discrimination for this category. Nevertheless, occasional confusion with neighboring heartbeat classes was observed, most likely because the limited number of available training examples restricted the ability of the model to learn the full range of morphological variability associated with fusion beats. This limitation is characteristic of highly imbalanced biomedical datasets and remains a common challenge in automated arrhythmia classification.

The Q class yielded no correct classifications in the present test partition; however, only seven Q beats were available for evaluation. Consequently, this result cannot provide a reliable estimate of the classifiers' intrinsic ability to recognize Q beats and should primarily be interpreted as a limitation of minority-class representation in the predefined patient-independent split. The heterogeneous composition of the Q category may additionally contribute to classification difficulty, but this cannot be established reliably from the present test support.

Overall, the confusion matrix demonstrates that most classification errors occurred between heartbeat classes exhibiting similar electrophysiological characteristics rather than between physiologically distinct categories. This observation confirms that the proposed feature set successfully captures major differences in cardiac electrical activity while highlighting the persistent difficulty of distinguishing morphologically overlapping heartbeat classes. Similar error patterns have been reported in recent patient-independent ECG classification studies, where supraventricular ectopic beats consistently represent the most challenging category because of their limited representation and strong morphological resemblance to normal sinus beats [10, 11, 16, 17].

From a clinical perspective, these findings are encouraging because the classifier achieved reliable recognition of normal and ventricular ectopic beats, which constitute the majority of routine ECG recordings and include clinically significant ventricular arrhythmias. Nevertheless, improving the discrimination of supraventricular ectopic beats remains an important objective for future research. Potential strategies include advanced feature engineering, cost-sensitive learning, synthetic minority oversampling, and hybrid machine-learning/deep-learning approaches specifically designed to enhance minority-class recognition while preserving the robustness of patient-independent evaluation.

G. Feature-Space Sensitivity Analysis

To evaluate the numerical stability of the proposed machine-learning framework, a supplementary feature-space sensitivity analysis was performed by introducing additive Gaussian perturbations into the extracted ECG feature vectors at different signal-to-noise ratio (SNR) levels. The corresponding classification accuracies obtained with the Random Forest and CatBoost classifiers are presented in Table VII and Fig. 7.

TABLE VII: SENSITIVITY OF RANDOM FOREST AND CATBOOST CLASSIFIERS TO FEATURE-SPACE PERTURBATIONS

SNR Level	Random Forest Accuracy (%)	CatBoost Accuracy (%)
Clean	89.34	88.85
30 dB	89.71	88.80
20 dB	90.93	88.57
10 dB	90.84	88.26
5 dB	90.28	88.10
0 dB	90.05	86.79

Classification accuracy obtained after introducing additive Gaussian perturbations at different signal-to-noise ratio (SNR) levels in the extracted feature space. Random Forest maintained stable performance across perturbation levels, whereas CatBoost exhibited a progressive decline under increasing perturbation intensity.

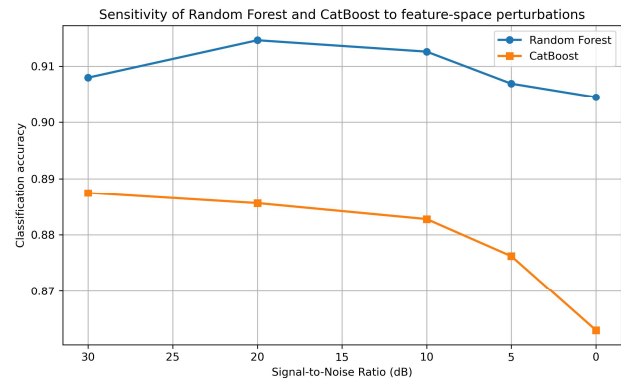


Fig. 7. Sensitivity of random forest and CatBoost classifiers to feature-space perturbations.

Classification accuracy obtained after the introduction of additive Gaussian perturbations at different Signal-to-Noise Ratio (SNR) levels in the extracted feature space. These experiments evaluate classifier sensitivity to feature perturbations rather than robustness to noise in raw ECG recordings and should therefore be interpreted accordingly.

The results demonstrate that the Random Forest classifier maintained remarkably stable performance across all perturbation levels. Classification accuracy varied only slightly, from 89.34% under clean conditions to approximately 90.05% at the highest perturbation level (0 dB). This limited variation indicates that the decision boundaries learned by the Random Forest model were only minimally affected by moderate changes in the extracted feature values. This limited variation indicates that the decision boundaries learned by the Random

Forest model were only minimally affected by controlled changes in the extracted feature values, suggesting high numerical stability within the engineered feature space.

In contrast, CatBoost exhibited a gradual reduction in performance as perturbation intensity increased. Classification accuracy decreased from 88.85% under clean conditions to 86.79% at 0 dB, indicating greater sensitivity to variations in the input features. Although the performance degradation remained relatively moderate, these findings suggest that CatBoost relies more strongly on precise feature values for decision making, whereas Random Forest benefits from the averaging effect of multiple decision trees, making it inherently more tolerant to feature perturbations.

These observations highlight an important characteristic of ensemble-learning algorithms. Because Random Forest aggregates predictions from a large number of independently constructed decision trees, local perturbations affecting individual features are less likely to substantially modify the final prediction. Conversely, boosting-based algorithms iteratively optimize residual errors, making their predictive performance potentially more sensitive to variations in feature distributions. The behavior observed in the present study is therefore consistent with the theoretical properties of these learning strategies.

It should be emphasized, however, that the present experiments were performed exclusively within the extracted feature space rather than on the original ECG waveforms. Consequently, the introduced perturbations simulate variations in the numerical feature values rather than realistic acquisition noise affecting ECG recordings. Therefore, these results should be interpreted as an evaluation of feature-space robustness and classifier stability, rather than as a direct assessment of resistance to physiological or instrumental ECG noise.

Nevertheless, the proposed sensitivity analysis provides additional evidence regarding the numerical stability of the developed classification framework under controlled perturbations of the extracted feature values. It should not be interpreted as evidence of robustness to noise affecting raw ECG recordings.

Evaluating classifier stability under controlled perturbations complements conventional performance metrics by assessing the consistency of model behavior under imperfect operating conditions. Such analyses are particularly relevant for biomedical engineering applications, where physiological variability, signal preprocessing, and measurement uncertainty may introduce moderate changes in extracted descriptors under practical signal-processing conditions.

Overall, the results indicate that the proposed feature-engineering approach generates a stable and informative representation of ECG signals, allowing reliable classification even when moderate perturbations are introduced into the feature space. The limited variation observed for the Random Forest classifier indicates comparatively high numerical stability under controlled feature-space perturbations. Future studies should separately evaluate robustness to realistic ECG

acquisition noise by introducing physiologically and instrumentally relevant noise directly into the raw ECG waveforms before feature extraction.

H. Overall Discussion

The present study demonstrates that classical machine-learning algorithms remain highly effective for automated ECG arrhythmia classification when combined with physiologically meaningful feature engineering and evaluated under a rigorous patient-independent inter-record validation framework. Rather than relying exclusively on increasingly complex deep-learning architectures [18–21], the proposed methodology emphasizes the integration of handcrafted morphological descriptors with temporal rhythm information, resulting in a compact and interpretable representation of cardiac electrical activity. The comparable performance achieved by Random Forest, XGBoost, and CatBoost further suggests that the quality of feature extraction and the rigor of the validation strategy play a more decisive role than algorithmic complexity alone in determining classification performance.

A major strength of this work lies in the adoption of patient-independent inter-record validation, which prevents information leakage between training and testing datasets. Unlike random beat-level partitioning, this protocol requires classifiers to generalize to ECG recordings originating from previously unseen patients, thereby providing a more realistic estimate of generalization performance on unseen patients. Recent systematic reviews continue to identify patient-independent validation as an essential prerequisite for fair comparison and reliable assessment of automated ECG classification systems [9–11].

Another important contribution is the emphasis placed on model interpretability. The feature-importance analysis demonstrated that temporal rhythm descriptors, particularly *RR_prev*, together with higher-order statistical measures such as kurtosis and skewness, were among the most influential predictors. Unlike many deep-learning models that provide limited insight into the decision-making process, the proposed framework enables direct interpretation of the physiological relevance of each feature, thereby improving model transparency and facilitating clinical acceptance. This aspect is increasingly recognized as a key requirement for trustworthy artificial intelligence in healthcare [13–15, 22].

The supplementary feature-space sensitivity analysis further complements the evaluation by assessing numerical stability under controlled perturbations of the extracted feature values. In particular, the Random Forest classifier maintained consistent performance across the evaluated perturbation levels. However, because perturbations were introduced after feature extraction rather than directly into the raw ECG waveforms, these findings should not be interpreted as evidence of robustness to physiological, electrode, motion, baseline-wander, or other acquisition-related ECG noise. Evaluation using realistic noise contamination applied

directly to raw ECG signals should therefore be addressed in future work.

Despite these encouraging findings, several limitations should be acknowledged. First, the study relied exclusively on the MIT-BIH Arrhythmia Database, which, although widely accepted as a benchmark dataset, may not fully represent the diversity of ECG recordings encountered in routine clinical practice.

Second, the marked class imbalance substantially affected minority-class evaluation. In particular, only seven Q beats were represented in the independent test partition, making the zero precision, recall, and F1-score obtained for this class statistically unreliable as an indicator of intrinsic classifier performance. Future investigations should consider patient-level cross-validation or repeated patient-independent evaluation strategies to obtain more reliable estimates for severely underrepresented heartbeat classes.

Third, the proposed framework was evaluated using handcrafted ECG descriptors only; therefore, it may not capture certain complex latent representations that can be learned directly from raw ECG signals by advanced deep-learning architectures.

Future research should therefore focus on validating the proposed methodology using independent multicenter ECG databases, investigating hybrid approaches that combine handcrafted descriptors with deep feature learning, and exploring advanced strategies for improving minority-class recognition while preserving patient-independent evaluation. Additional work integrating explainable artificial intelligence techniques and real-time implementation on embedded or wearable ECG monitoring systems may further facilitate the translation of automated arrhythmia classification into routine clinical practice and telemedicine applications, consistent with recent developments in intelligent biomedical engineering and Internet of Medical Things technologies [23–25].

In summary, the findings of the present study indicate that interpretable feature-based machine-learning models can provide a valuable complementary approach to deep-learning methods, particularly when model transparency and computational efficiency are important considerations. Because deep-learning models were not directly evaluated under the same patient-independent protocol in this study, direct comparative assessment should be addressed in future work.

When supported by rigorous patient-independent validation, comprehensive performance analysis, and physiologically meaningful feature engineering, such models can provide reliable and transparent computational tools with potential relevance for future clinical decision-support applications.

IV. CONCLUSION

This study presented a feature-based machine-learning framework for automated ECG arrhythmia classification using the MIT-BIH Arrhythmia Database under a rigorous patient-independent inter-record validation protocol. By combining physiologically meaningful

morphological descriptors with RR-interval features, the proposed approach achieved reliable heartbeat classification while maintaining a high level of interpretability and computational efficiency.

Among the evaluated classifiers, Random Forest achieved the highest overall classification accuracy, whereas CatBoost provided the best balance between precision and F1-score. Receiver operating characteristic analysis demonstrated excellent discrimination for ventricular ectopic beats, while supraventricular ectopic beats remained the most challenging category because of their morphological similarity to normal beats and the severe class imbalance inherent to the dataset. Feature-importance analysis further revealed that temporal rhythm descriptors, particularly the previous RR interval, together with higher-order statistical descriptors such as kurtosis and skewness, played a dominant role in the classification process. In addition, the feature-space sensitivity analysis showed that the Random Forest classifier maintained stable performance under moderate perturbations, supporting the robustness of the proposed feature representation.

Beyond its predictive performance, the principal contribution of this work lies in the combination of rigorous patient-independent validation, interpretable feature engineering, feature-importance analysis, and feature-space sensitivity analysis within a unified machine-learning framework. The proposed methodology demonstrates that carefully engineered ECG descriptors can provide accurate, transparent, and computationally efficient arrhythmia classification while offering direct insight into the physiological variables contributing to automated decision making. These characteristics position feature-based machine learning as a complementary approach to deep-learning methods rather than as a demonstrably superior alternative.

Nevertheless, several limitations should be acknowledged. The study was conducted using a single publicly available database, and the pronounced imbalance among heartbeat classes, particularly for supraventricular ectopic and fusion beats, inevitably influenced classification performance.

Furthermore, the sensitivity analysis assessed numerical stability under controlled perturbations of the extracted feature values and did not evaluate robustness to noise affecting the original ECG waveforms. Consequently, future investigations should evaluate the proposed framework using realistic ECG noise introduced directly into raw signals before feature extraction, in addition to validation on independent multicenter datasets.

Future research should focus on external validation using diverse clinical databases, direct comparison with deep-learning architectures evaluated under the same patient-independent experimental protocol, the integration of explainable artificial intelligence techniques, and the development of hybrid frameworks combining handcrafted ECG descriptors with deep feature-learning approaches.

Such strategies may improve minority-class recognition while preserving model transparency and

computational efficiency. In addition, implementation on embedded systems, wearable ECG devices, and Internet of Medical Things platforms represents a promising direction for evaluating the potential applicability of the proposed methodology to future real-time and remote ECG-monitoring systems.

Overall, the present findings demonstrate that interpretable feature-based machine-learning models remain a valuable solution for automated ECG arrhythmia classification. When combined with rigorous patient-independent evaluation and physiologically meaningful ECG descriptors, these approaches provide reliable and transparent computational tools with potential relevance for future clinical decision-support applications.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

AUTHOR CONTRIBUTIONS

Leila Sahraoui: Conceptualization, methodology, software development, data preprocessing, feature extraction, formal analysis, data curation, visualization, and writing of the original draft. Abdelaziz Frih: Methodology, software validation, data analysis, interpretation of results, manuscript review, and editing. Hacène Frih: Study conception and design, supervision, methodology, interpretation of findings, critical revision of the manuscript for important intellectual content, project administration, and final approval of the manuscript. All authors read and approved the final version of the manuscript and agree to be accountable for all aspects of the work.

ACKNOWLEDGMENT

The authors would like to thank the developers and contributors of the MIT-BIH Arrhythmia Database and PhysioNet for providing open access to this valuable research resource. The authors also acknowledge Badji Mokhtar University, Annaba, Algeria, for supporting the research environment in which this study was conducted.

REFERENCES

- [1] Timmis, P. Vardas, N. Townsend, A. Torbica, H. A. Katus, D. Smedt *et al.*, "European Society of Cardiology: The 2023 Atlas of Cardiovascular Disease Statistics," *Eur. Heart J.*, vol. 45, no. 38, pp. 4019–4106, 2024. doi: 10.1093/eurheartj/ehae420
- [2] Y. Ansari, O. Mourad, K. Qaraqe, and E. Serpedin, "Deep Learning for ECG Arrhythmia Detection and Classification: An Overview of Progress for Period 2017–2023," *Front. Physiol.*, vol. 14, 1246746, 2023. doi: 10.3389/fphys.2023.1246746
- [3] N. Katal, S. Gupta, P. Verma, and B. Sharma, "Deep-Learning-Based Arrhythmia Detection Using ECG Signals: A Comparative Study and Performance Evaluation," *Diagnostics*, vol. 13, no. 24, 3605, 2023. doi: 10.3390/diagnostics13243605
- [4] A. J. Prakash, K. Chandrasekaran *et al.*, "Advances in Machine and Deep Learning for ECG Beat Classification," *Front. Digit. Health*, vol. 7, 1649923, 2025. doi: 10.3389/fdgth.2025.1649923
- [5] Q. Xiao, K. Lee, S. A. Mokhtar, I. Ismail, A. L. B. M. Pauzi, Q. Zhang *et al.*, "Deep Learning-Based ECG Arrhythmia Classification: A Systematic Review," *Appl. Sci.*, vol. 13, no. 8, 4964, 2023. doi: 10.3390/app13084964
- [6] H. Velandia, J. Rueda-Arias *et al.*, "Systematic Review of Artificial Intelligence and Electrocardiography for Cardiovascular Disease Detection," *Bioengineering*, vol. 12, no. 11, 1248, 2025. doi: 10.3390/bioengineering12111248
- [7] S. K. Jamage, R. Y. Mali, and V. V. Shete, "Optimizing Myocardial Infarction Prediction Performance Using Advanced Machine Learning Techniques," *Int. J. Elect. Electron. Eng. Telecommun.*, vol. 14, no. 5, pp. 282–295, 2025. doi: 10.18178/ijeetc.14.5.282-295
- [8] X. Bai, X. Dong, Y. Li, R. Liu, and H. Zhang, "A Hybrid Deep Learning Network for Automatic Diagnosis of Cardiac Arrhythmia Based on 12-Lead ECG," *Sci. Rep.*, vol. 14, 24441, 2024. doi: 10.1038/s41598-024-75531-w
- [9] P. de Chazal, M. O'Dwyer, and R. B. Reilly, "Automatic Classification of Heartbeats Using ECG Morphology and Heartbeat Interval Features," *IEEE Trans. Biomed. Eng.*, vol. 51, no. 7, pp. 1196–1206, Jul. 2004. doi: 10.1109/TBME.2004.827359
- [10] C. Zhou, X. Sun, Y. Wang, and Y. Zhang, "Inter-Patient ECG Heartbeat Classification for Arrhythmia Using a Weighted Capsule Network Combined with Sequence-to-Sequence Learning Under Class Imbalance Conditions," *Front. Physiol.*, vol. 14, 1247587, 2023. doi: 10.3389/fphys.2023.1247587
- [11] G. A. L. Silva, P. H. L. Silva, G. J. P. Moreira, V. L. S. Freitas, J. C. Gertrudes, and E. J. S. Luz, "A Systematic Review of ECG Arrhythmia Classification: Adherence to Standards, Fair Evaluation, and Embedded Feasibility," arXiv, Preprint, arXiv:2503.07276, 2025.
- [12] A. Kirkbas and A. Kizilkaya, "Automated ECG Arrhythmia Classification Using Feature Images with Common Matrix Approach-Based Classifier," *Sensors*, vol. 25, no. 4, 1220, 2025. doi: 10.3390/s25041220
- [13] A. E. M. Atwa, M. Elgendi *et al.*, "Interpretable Deep Learning Models for Arrhythmia Classification," *Diagnostics*, vol. 15, no. 15, 1950, 2025. doi: 10.3390/diagnostics15151950
- [14] Z. Al-Odat, "Optimized Distributed Gradient Boosting with Explainable Artificial Intelligence for Real-Time Ransomware Detection in Networked Environments," *Int. J. Elect. Electron. Eng. Telecommun.*, vol. 14, no. 4, pp. 233–242, 2025. doi: 10.18178/ijeetc.14.4.233-242
- [15] M. A. Talukder, M. M. Islam *et al.*, "An Explainable Deep Learning Framework for Trustworthy Arrhythmia Detection from ECG Signals," *Sci. Rep.*, vol. 15, 39496, 2025. doi: 10.1038/s41598-025-22986-0
- [16] Y. Jeong, J. Kim, S. Lee, and H. Park, "Enhancing Inter-Patient Performance for Arrhythmia Classification Using Domain-Generalization Strategies and Deep Learning Models," *Appl. Sci.*, vol. 14, no. 16, 7227, 2024. doi: 10.3390/app14167227
- [17] S. Mousavi, F. Afghah, and U. R. Acharya, "Inter-Patient ECG Heartbeat Classification Using Convolutional and Recurrent Neural Networks," *IEEE Access*, vol. 8, pp. 116868–116879, 2020. doi: 10.1109/ACCESS.2020.3003879
- [18] A. H. Ribeiro, M. H. Ribeiro, G. M. M. Paixão, D. M. Oliveira, P. R. Gomes, J. A. Canazart, M. P. Ferreira, C. R. Andersson, P. W. Macfarlane, M. B. Wagner, N. Samesima, and A. L. P. Ribeiro, "Automatic Diagnosis of the 12-Lead ECG Using a Deep Neural Network," *Nat. Commun.*, vol. 11, 1760, 2020. doi: 10.1038/s41467-020-15432-4
- [19] A. Shah, M. Khan *et al.*, "Electrocardiogram Analysis for Cardiac Arrhythmia Classification Using Self-Attention Autoencoder Models," *Sci. Rep.*, vol. 15, 9230, 2025. doi: 10.1038/s41598-025-93906-5
- [20] M. R. Baidillah, P. Busono, I. M. Astawa, S. Karim, R. Febryarto, I. P. A. Yogiswara *et al.*, "Hybrid Kolmogorov-Arnold and Convolutional Neural Network Model for Single-Lead Electrocardiogram Classification," *TELKOMNIKA Telecommun. Comput. Electron. Control*, vol. 23, no. 5, pp. 1342–1352, 2025. doi: 10.12928/TELKOMNIKA.v23i5.26735
- [21] E. Oporto, A. Morales *et al.*, "Challenges in the Classification of Cardiac Arrhythmias and Ischemia Using Deep Learning and ECG Signals," *Diagnostics*, vol. 16, no. 1, 161, 2026. doi: 10.3390/diagnostics16010161
- [22] M. Lee, J. Lim, and J. Kim, "ECG-GraphNet: Advanced Arrhythmia Classification Based on Graph Convolutional Networks," *Heart Rhythm O2*, vol. 6, no. 8, pp. 1199–1211, 2025. doi: 10.1016/j.hroo.2025.05.012

- [23] H. M. Tun, "Photoplethysmography (PPG) Scheming System Based on Finite Impulse Response (FIR) Filter Design in Biomedical Applications," *Int. J. Elect. Electron. Eng. Telecommun.*, vol. 10, no. 4, pp. 272–282, 2021. doi: 10.18178/ijeetc.10.4.272-282
- [24] L. S. Ahmed and A. I. Siddiq, "A Comprehensive Review of the Internet of Medical Things in Healthcare," *Int. J. Elect. Electron. Eng. Telecommun.*, vol. 13, no. 6, pp. 415–426, 2024. doi: 10.18178/ijeetc.13.6.415-426
- [25] A. A. M. Bulbul, S. Ahmed *et al.*, "Deep Learning for ECG-Based Cardiovascular Disease Diagnosis: A Systematic Review," *Artif. Intell. Rev.*, vol. 59, 136, 2026. doi: 10.1007/s10462-025-11479-1

Copyright © 2026 by the authors. This is an open access article distributed under the Creative Commons Attribution License which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited ([CC BY 4.0](https://creativecommons.org/licenses/by/4.0/)).

Leila Sahraoui was born in Annaba, Algeria, on September 14, 1972. She received the state engineer degree in electronics, specializing in industrial control, from Badji Mokhtar University of Annaba, Annaba, Algeria, in 1998; the magister degree in microwaves for Telecommunications from the University of Batna, Batna, Algeria, in 2008; and the doctorate ès sciences degree in telecommunications from Badji Mokhtar University of Annaba, Annaba, Algeria, in 2020. She is currently a faculty member in the Department of Electronics, Faculty of Technology, Badji Mokhtar University of Annaba, Annaba, Algeria. Her teaching and research activities are mainly related to telecommunications, communication systems, and signal processing. Her current research interests include digital communications, signal and image processing, information and communication technologies, artificial intelligence, machine learning, and their applications in health

and environmental sciences. She is a member of the ICT, Health and Environment research team at Badji Mokhtar University of Annaba.

Abdelaziz Frih was born in Algeria. He received. He is affiliated with the Department of Biology, Faculty of Sciences, Badji Mokhtar University of Annaba, Annaba, Algeria, and the Environmental Biomonitoring Laboratory. His research activities focus on ecology, environmental biomonitoring, biodiversity, freshwater turtle biology and conservation, and the effects of environmental factors on wildlife. His current research interests include conservation biology, freshwater ecology, ecotoxicology, environmental pollution, biodiversity, and the ecology of Mediterranean and North African ecosystems.

Hacène Frih was born in Annaba, Algeria, on November 12, 1970. He received the DES degree in animal biology from Badji Mokhtar University of Annaba, Annaba, Algeria, in 1996; the magister degree in biology from Badji Mokhtar University of Annaba, Annaba, Algeria, in 1999; the doctorate ès sciences degree in biology from Badji Mokhtar University of Annaba, Annaba, Algeria, in 2007; and the Habilitation to Direct Research (HDR) from Badji Mokhtar University of Annaba, Annaba, Algeria, in 2009. He is currently a professor in the Department of Biology, Faculty of Sciences, Badji Mokhtar University of Annaba, Annaba, Algeria, where he is involved in undergraduate, postgraduate, and doctoral education and research. He is responsible for doctoral training in physiology and pathophysiology and leads the "Stress and Cellular Activity" research team. His scientific work has focused on physiology and pathophysiology, neuro-immuno-endocrinology, neurophysiology of stress, neuroscience, immunology, and psychopharmacology. His current research interests also include biomedical data analysis, machine learning, artificial intelligence in health sciences, physiological signal processing, and predictive modeling.