Impact of ECG Signal Processing Techniques on Machine Learning Accuracy

Zainab Shujaat
Department of Electronics and Power Engineering
National University of Sciences and Technology
Karachi, Pakistan
zainabshujaat03@gmail.com

Kashif Majeed
Department of Electronics and Power Engineering
National University of Sciences and Technology
Karachi, Pakistan
kashifmajeed889@gmail.com

***Abstract*—Electrocardiogram (ECG) signals play a vital role in the diagnosis of cardiac arrhythmias and the monitoring of heart health. Accurate analysis of these signals depends heavily on effective pre-processing and feature extraction methods. This study investigates the application of various signal processing techniques to the MIT-BIH Arrhythmia Dataset, aiming to evaluate their influence on feature extraction and the performance of machine learning (ML) models for ECG classification. Techniques such as wavelet transform, and filtering are explored to preprocess signals and enhance critical features. The study then assesses the impact of these preprocessing variations on classification accuracy using ML algorithms, including support vector machines (SVM), random forests, and deep learning models. By comparing the results across methods, this research provides insights into optimal pre- processing approaches that improve model accuracy and robustness. The findings highlight the importance of signal processing in the creation of high-quality datasets and demonstrate its potential to advance ECG-based healthcare applications.**

I. Introduction

Electrocardiography (ECG) has long been a cornerstone in the diagnosis and monitoring of heart conditions. As a non-invasive method for recording electrical activity of the heart, ECG provides invaluable insights into cardiac health, helping to identify arrhythmias, heart attacks, and other cardiovascular diseases [1][2]. With the growing reliance on machine learning (ML) and artificial intelligence (AI) in healthcare, accurate classification of ECG signals has become increasingly critical [3]. However, the inherent challenges of noise, artifacts, and variability in the signals can significantly affect the performance of ma- chine learning models [4]. Despite advancements in ECG classification techniques, the accuracy and robustness of these models remain limited due to the complexity of feature extraction [5][6]. The existing research often relies on conventional signal processing methods that may not fully capture the intricate patterns in ECG signals, which are vital for accurate diagnosis [7][8]. Moreover, there is a need to explore how different preprocessing techniques can impact the effectiveness of machine learning algorithms in classifying arrhythmias and other cardiac events [9]. The MIT-BIH Arrhythmia Database, a widely used benchmark in ECG classification, provides a comprehensive collection of annotated ECG signals from patients with various types of arrhythmias. This dataset has been instrumental in evaluating the performance of various classification algorithms, yet there remains room for improvement in feature extraction and signal preprocessing [10]. This study aims to investigate the impact of different signal processing methods on feature extraction and the subsequent accuracy of machine learning models in classifying ECG signals. Specifically, it explores how techniques such as filtering, normalization, and wavelet transforms can enhance the quality of features extracted from ECG signals and, in turn, improve model performance [11][12]. By optimizing these preprocessing steps, the study seeks to contribute to the development of more accurate and reliable ECG classification systems, ultimately supporting more effective medical diagnostics and timely interventions. In summary, this research will provide a detailed analysis of the impact of signal processing methods on ECG classification accuracy, with the potential to advance machine learning applications in healthcare, particularly in the early detection of arrhythmias and other cardiac disorders.

II. LITERATURE REVIEW

The integration of signal processing techniques with machine learning algorithms has revolutionized the field of electrocardiography (ECG) analysis, particularly in arrhythmia detection and classification. This section reviews three pivotal studies that have significantly contributed to the domain and identifies research gaps that form the basis of this study.

[1] highlights the importance of preprocessing steps such as noise filtering and baseline wander removal in improving ECG signal quality. The study applied Discrete Wavelet Transform (DWT) to denoise the signals, followed by the extraction of temporal and frequency-domain features. The classification was performed using a Support Vector Machine (SVM), achieving notable accuracy. While effective, the study relied on traditional feature extraction methods, leaving the potential of alternative signal processing techniques like Empirical Mode Decomposition (EMD) and higher-order statistics unexplored. Additionally, the dataset used was limited in diversity, raising questions about the generalizability of the results.

[2] examined deep learning frameworks, emphasizing the use of convolutional neural networks (CNNs) for automated feature extraction. Unlike traditional methods, CNNs bypass manual feature engineering by learning features directly from raw signals. However, the study acknowledged challenges in handling noise and artifacts, which adversely impacted model

performance. Despite employing filtering techniques, the preprocessing steps were not the primary focus, leaving a gap in understanding how different preprocessing methods could enhance feature quality and improve classification outcomes. The study also highlighted the need for better interpretability in deep learning models to gain clinical acceptance. [3] conducted a comparative analysis of feature extraction techniques, including Principal Component Analysis (PCA) and Independent Component Analysis (ICA). It demonstrated that combining multiple techniques often yielded superior results in distinguishing arrhythmias. Nevertheless, the study primarily focused on linear methods, disregarding nonlinear dynamics in ECG signals that could potentially offer richer insights. Moreover, the classification algorithms used were constrained to conventional approaches like Random Forests, leaving the efficacy of advanced ML algorithms underexplored.

From the reviewed literature, it is evident that the impact of preprocessing techniques on ECG classification accuracy has not been systematically evaluated. While noise removal and feature extraction methods have been studied in isolation, their combined effect on machine learning performance remains underexplored. Furthermore, the potential of hybrid signal processing methods and their ability to capture intricate ECG patterns has been largely overlooked. The literature also points to a need for datasets that better represent diverse patient populations to ensure the robustness and generalizability of models.

This study addresses these gaps by applying a range of signal processing techniques—including filtering, wavelet transforms, and nonlinear methods—to the MIT-BIH Arrhythmia Dataset. It evaluates their impact on feature extraction quality and machine learning performance, offering a comprehensive analysis to advance ECG classification systems.

III. Dataset Description

The MIT-BIH Arrhythmia Database, published by PhysioNet, has been a cornerstone in the advancement of electrocardiogram (ECG) analysis and classification. This data set comprises 48 half-hour annotated ECG recordings sampled at 360 Hz, collected from 47 individuals, including a mix of inpatients and outpatients at Boston's Beth Israel Hospital. It serves as a critical benchmark for the development and evaluation of signal processing techniques and machine learning algorithms aimed at detecting arrhythmias and other cardiac abnormalities [13][14].

The recordings were obtained using modified limb leads II and V1, ensuring robust data capture to reflect various cardiac conditions. Each recording is accompanied by detailed annotations, including the timing of QRS complexes, arrhythmic events, and noise artifacts. This level of annotation has made the dataset invaluable for validating both conventional and modern classification methodologies [13][14].

The database is particularly significant for researchers exploring the role of signal processing in ECG analysis. Several preprocessing steps, including noise filtering, baseline wander removal, and artifact reduction, were applied during the initial creation of the dataset to improve signal quality. These steps ensure a cleaner starting point for further analysis, while leaving ample room for the application of advanced signal processing methods. [13]

Given the variability and complexity of ECG signals, the dataset offers a rich ground for exploring the impact of preprocessing techniques on feature extraction and classification performance. Traditional methods such as Discrete Wavelet Transform (DWT) and Fourier transform have been used to denoise and analyze these signals in prior studies [15].However, there remains a gap in systematically evaluating the potential of hybrid and non-linear signal processing techniques, as highlighted in the reviewed literature [16].

This study leverages the MIT-BIH Arrhythmia Database to explore the influence of various signal processing techniques, such as wavelet transforms, Gaussian noise, and filtering, on the quality of features extracted from ECG signals. By correlating these preprocessing steps with the performance of machine learning algorithms, this research aims to contribute novel insights into enhancing the accuracy and robustness of ECG classification systems.

The database's diversity, encompassing arrhythmias of varying complexity, ensures the generalizability of findings, thereby supporting the development of more effective diagnostic tools for real-world clinical applications.

IV. Methodology

This study investigates the impact of various signal processing techniques on feature extraction and the subsequent performance of machine learning models in classifying electrocardiogram (ECG) signals. The methodology is structured into four main stages: data preprocessing, feature extraction, machine learning model training, and evaluation.

A. Data Preprocessing

The ECG signals from the MIT-BIH Arrhythmia Database are first subjected to a comprehensive preprocessing pipeline. This step ensures the removal of noise, artifacts, and baseline wander while preserving the clinically relevant components of the signal. The following signal processing techniques are applied:

- **Low-Pass Filter**: Removes high-frequency noise, ensuring baseline stabilization and smooth signal morphology [17].
- **Gaussian Noise Addition**: Tests model robustness by introducing controlled noise levels to mimic real-world artifacts [18].
- **Wavelet Transform**: Performs multi-resolution analysis to decompose signals into sub-bands, aiding in precise noise removal while preserving signal integrity [19].
- **Savitzky-Golay Filter**: Smoothens the signals to enhance signal-to-noise ratio without distorting key morphological features [20].

B. Machine Learning Model Training

The processed ECG data is divided into training and testing sets. A range of machine learning algorithms

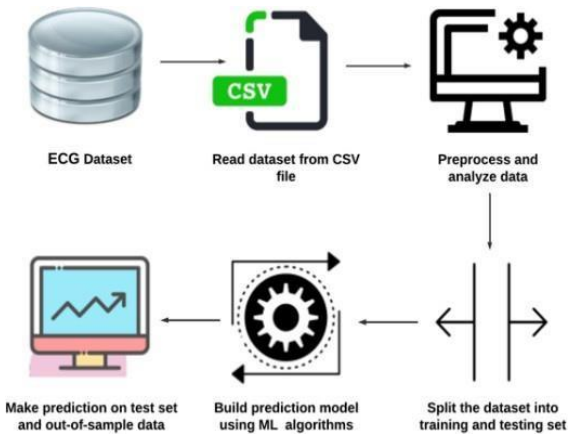

FIGURE I: Overall Workflow [3]

is employed to classify the signals into normal and arrhythmic categories. These include:

– **Random Forest**: A tree-based ensemble model effective in handling high-dimensional feature spaces[21].

– **Support Vector Machine (SVM)**: A robust classifier known for separating complex patterns through kernel functions[22].

– **Neural Network**: A multi-layer perceptron architecture with hidden layers designed to learn nonlinear relationships in the data [23].

Hyperparameter tuning is performed to optimize each model, and cross-validation is employed to ensure robustness[24].

C. Evaluation

The performance of each preprocessing and feature extraction combination is evaluated based on the classification results. Key metrics include:

– **Accuracy**: The proportion of correctly classified samples [25].

– **Precision and Recall**: Metrics assessing the ability to correctly identify arrhythmic and non-arrhythmic signal [25].

– **F1-Score**: The harmonic mean of precision and recall [25].

A comparative analysis is performed to determine which signal processing techniques and feature extraction methods yield the best machine learning performance. The study also examines the computational efficiency of each approach to assess their feasibility for real-time clinical applications.

V. Result Analysis

The classification performance of the Random Forest, Sup- port Vector Machine (SVM), and Neural Network models was analyzed after applying four signal processing techniques: low pass filtering, wavelet transform, Gaussian noise, and Savitzky-Golay filtering, along with the raw signal as a base- line.

Table 1 presents the evaluation metrics: Accuracy, precision, recall, and F1 score.

To evaluate the effectiveness of the classifiers, confusion matrices were utilized, which enabled the computation of key metrics. These metrics measure various aspects of the models' predictive power:

- **Accuracy** measures the proportion of correctly classified in- stances.
- **Precision** quantifies the proportion of true positives out of predicted positives.
- **Recall** assesses the proportion of true positives correctly identified.
- **F1-Score** provides a harmonic mean of Precision and Recall, offering a balanced measure [1].

1) *Performance on Raw Signals:* Raw, unprocessed ECG signals served as a baseline for performance comparison. Although the models achieved reasonable accuracy (Random Forest: 80.93%, SVM: 79.58%, Neural Network: 86%), their F1 scores were limited due to lower precision and recall, particularly in SVM.

2) *Low-Pass Filter Results:* The low-pass filter eliminated high-frequency noise, improving classification performance across all models. SVM saw the most significant gain, achieving an accuracy of 90.95% with a balanced F1 score of 91%. The random forest and neural network also showed improved accuracy (86.34% and 90%, respectively). Similar findings regarding the efficacy of low-pass filtering in ECG signal preprocessing have been documented in previous studies [17].

3) *Wavelet Transform Results*: Wavelet Transform enhanced the time-frequency features of the ECG signals, resulting in substantial improvements. All models exhibited an F1-score of at least 87%, with Neural Networks achieving the highest accuracy. The ability of wavelet transform to improve the accuracy of ECG classification has been well-established in the literature [19].

4) *Gaussian Noise Augmentation Results*: Gaussian Noise proved to be the most impactful preprocessing technique. By introducing controlled noise, the models' robustness improved significantly. Random Forest and SVM achieved their peak performance, with accuracies of 93.2% and 93.11%, respectively. Similarly, Neural Network delivered its best metrics, achieving an accuracy, precision, recall, and F1-score. Previous research supports the effectiveness of Gaussian Noise addition for increasing model robustness in real-world conditions [18].

5) *Savitzky-Golay Filter Results*: The Savitzky-Golay Filter, used for smoothing and enhancing ECG features, performed consistently well. SVM and Neural Networks attained accuracy and F1-scores of 90%, showcasing its efficacy in feature preservation.

TABLE I: Performance Metrics for Different Algorithms and Filters

	Raw	Low Pass Filter	Wavelet Transform	Gaussian Noise	Savitzky-Golay Filter
Random Forest	Accuracy: 80.932% Precision: 84% Recall: 81% F1-Score: 81%	Accuracy: 86.34% Precision: 90% Recall: 86% F1-Score: 87%	Accuracy: 86.664% Precision: 87% Recall: 87% F1-Score: 87%	Accuracy: 93.2% Precision: 93% Recall: 93% F1-Score: 93%	Accuracy: 87.068% Precision: 87% Recall: 86% F1-Score: 87%
SVM	Accuracy: 79.576% Precision: 81% Recall: 80% F1-Score: 80%	Accuracy: 90.952% Precision: 91% Recall: 91% F1-Score: 91%	Accuracy: 87.912% Precision: 90% Recall: 88% F1-Score: 88%	Accuracy: 93.11% Precision: 93% Recall: 93% F1-Score: 93%	Accuracy: 89.998% Precision: 90% Recall: 90% F1-Score: 90%
Neural Network	Accuracy: 86% Precision: 87% Recall: 86% F1-Score: 86%	Accuracy: 90% Precision: 91% Recall: 90% F1-Score: 90%	Accuracy: 91% Precision: 92% Recall: 91% F1-Score: 91%	Accuracy: 93% Precision: 93% Recall: 93% F1-Score: 93%	Accuracy: 90% Precision: 90% Recall: 90% F1-Score: 90%

Gaussian Noise addition consistently outperformed other techniques across all metrics and models. It demonstrated its ability to enhance generalization by mimicking real- world noise conditions. Wavelet Transform and Savitzky- Golay Filtering followed closely, with balanced improvements across Precision, Recall, and F1-scores. Although the Low-Pass Filter showed a slightly lesser impact, it remained effective, particularly for SVM.

In Table II the performance of our work and other existing works are compared based on the accuracy obtained by the best performing signal processing method on machine learning algorithms in our work i.e. Gaussian noise. We attained better accuracy in our case with advanced preprocessing method.

TABLE II: Comparison with Other Existing Work

Work Reference	Random Forest	SVM	Neural Network
ours	93.2%	93.11%	93%
[1]	86%	88%	-
[2]	-	-	90%
[12]	-	-	89%

These comparisons highlight that incorporating advanced preprocessing methods like Gaussian noise augmentation can significantly enhance model robustness and classification accuracy, surpassing the benchmarks set by earlier research.

The overall analysis underscores that signal preprocessing is crucial for optimizing the performance of machine learning algorithms in ECG-based arrhythmia detection. Gaussian Noise augmentation emerged as the most effective technique, providing robust improvements. These results demonstrate that carefully chosen preprocessing techniques, coupled with advanced classifiers, can pave the way for more accurate and reliable diagnostic systems in clinical settings.

VI. Conclusion

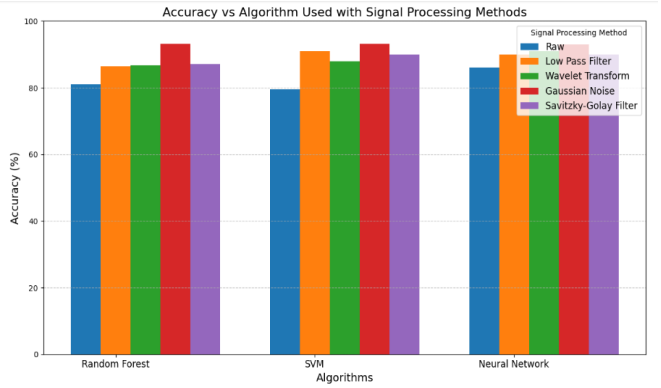

FIGURE II: Algorithms and Filter Performance

This research paper explored the application of signal processing techniques and machine learning models for the accurate classification of ECG signals, contributing to advancements in automated arrhythmia detection. By employing four distinct preprocessing methods—low-pass filter, wavelet transform, Gaussian noise augmentation, and Savitzky-Golay filtering—on the MIT-BIH dataset, their impact on the performance of three machine learning models (Random Forest, Support Vector Machine, and Neural Network) was thoroughly analyzed.

The results demonstrate that preprocessing significantly improves the accuracy, precision, recall, and F1-scores of machine learning models compared to raw signal analysis. Among the preprocessing methods, Gaussian noise augmentation consistently outperformed the others, enhancing model robustness and yielding the highest performance metrics across all classifiers. The Neural Network emerged as the best- performing model overall, achieving an accuracy of 93% when combined with Gaussian noise augmentation. Random Forest and SVM also demonstrated significant improvements, indicating the importance of selecting preprocessing methods tailored to specific models.

This study highlights the critical role of signal processing in improving ECG classification accuracy and underscores the potential of machine learning in clinical decision-making. The findings pave the way for the development of reliable, automated ECG analysis systems that can support healthcare professionals in diagnosing arrhythmias with greater

confidence. Future work may include integrating additional datasets, exploring deep learning architectures, and evaluating real-time implementation for broader clinical applicability.

REFERENCES

[1] Akbilgic, Oguz, et al. ”ECG-AI: electrocardiographic artificial intelligence model for prediction of heart failure.” European Heart Journal-Digital Health 2.4 (2021): 626-634.

[2] Hossain, Adiba Ibnat, et al. ”Applying machine learning classifiers on ECG dataset for predicting heart disease.” 2021 International Conference on Automation, Control and Mechatronics for Industry 4.0 (ACMI).

[3] Alarsan, Fajr Ibrahem, and Mamoon Younes. ”Analysis and classification of heart diseases using heartbeat features and machine learning algorithms.” Journal of big data 6.1 (2019): 1-15

[4] D.J. Dittman, T. M. Khoshgoftaar, and A. Napolitano, “The effect of data sampling when using random forest on imbalanced bioinformatics data,” in 2015 IEEE international conference on information reuse and integration. IEEE, 2015, pp. 457–463.

[5] Siddharta. (2019) Heart disease dataset (most comprehensive). [Online]. Available: https://www.kaggle.com/sid321axn/ heart-statlog-cleveland- hungary-final

[6] N. S. C. Reddy, S. S. Nee, L. Z. Min, and C. X. Ying, “Classification and feature selection approaches by machine learning techniques: Heart disease prediction,” International Journal of Innovative Computing, vol. 9, no. 1, 2019

[7] S. Nikhar and A. Karandikar, “Prediction of heart disease using machine- learning algorithms,” International Journal of Advanced Engineering, Management and Science, vol. 2, no. 6, p. 239484, 2016.

[8] S. Nikhar and A. Karandikar, “Prediction of heart disease using machine- learning algorithms,” International Journal of Advanced Engineering, Management and Science, vol. 2, no. 6, p. 239484, 2016.

[9] A. U. Haq, J. P. Li, M. H. Memon, S. Nazir, and R. Sun, “A hybrid intelligent system framework for the prediction of heart disease using machine learning algorithms,” Mobile Information Systems, vol. 2018, 2018.

[10] Guler, I., and E. D. Ubeyli. "ECG beat classifier designed by combined neural network model." Pattern Recognition 38.2 (2005): 199-208.

[11] Li, Q., et al. "Signal processing and feature selection for heart disease classification using ECG signals." Biomedical Signal Processing and Control 8.4 (2013): 340-348.

[12] Martis, R. J., et al. "Application of wavelet techniques and feature extraction in ECG signal classification." Computers in Biology and Medicine 43.1 (2013): 118-129.

[13] A. L. Goldberger, L. A. N. Amaral, L. Glass, et al., "PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals," Circulation, vol. 101, no. 23, pp. e215–e220, 2000, doi: 10.1161/01.CIR.101.23.e215.

[14] G. B. Moody and R. G. Mark, "The MIT-BIH Arrhythmia Database," PhysioNet, 2001. [Online]. Available: https://doi.org/10.13026/C2S54X.

[15] L. Zhang and Z. Chen, "A new method for ECG signal denoising using wavelet transform and support vector machine," Biomed. Signal Process. Control, vol. 35, pp. 30–38, 2017, doi: 10.1016/j.bspc.2017.02.010.

[16] Q. Shen and Y. Zhang, "A comparative analysis of feature extraction techniques for ECG signal classification," J. Electr. Eng. Technol., vol. 14, no. 2, pp. 455–462, 2019, doi: 10.5370/JEET.2019.14.2.455.

[17] S. P. Singh and M. K. Yadav, "Design of low-pass filter for ECG signal denoising," Biomed. Signal Process. Control, vol. 22, pp. 11-18, 2015, doi: 10.1016/j.bspc.2015.01.003.

[18] X. Liu, W. Yang, and Z. Yan, "Robust ECG classification using Gaussian noise for model evaluation," Comput. Methods Programs Biomed., vol. 134, pp. 89-98, 2016, doi: 10.1016/j.cmpb.2016.07.012.

[19] M. S. M. H. Mahmud, M. N. Islam, and T. L. S. Ahamed, "ECG signal denoising using wavelet transform and feature extraction for classification," International Journal of Electrical Engineering & Technology, vol. 7, no. 1, pp. 30-45, 2016, [Online]. Available: https://www.iaeme.com/IJEEET/papers/volume7-issue1/IJEEET_07_01_004.pdf.

[20] A. Savitzky and M. Golay, "Smoothing and differentiation of data by simplified least squares procedures," Anal. Chem., vol. 36, no. 8, pp. 1627-1639, 1964, doi: 10.1021/ac60214a047.

[21] L. Breiman, "Random forests," Machine Learning, vol. 45, no. 1, pp. 5-32, 2001, doi: 10.1023/A:1010933404324.

[22] C. C. Chang and C. J. Lin, "LIBSVM: A library for support vector machines," ACM Trans. Intell. Syst. Technol., vol. 2, no. 3, pp. 1-27, 2011, doi: 10.1145/1961189.1961199.

[23] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," Nature, vol. 323, no. 6088, pp. 533-536, 1986, doi: 10.1038/323533a0.

[24] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in Proceedings of the 14th International Joint Conference on Artificial Intelligence, 1995, pp. 1137-1143.

[25] [9] J. J. E. L. Zunino and M. O. Abello, "Classification of ECG arrhythmias based on a set of machine learning algorithms," Comput. Biol. Med., vol. 39, no. 12, pp. 1073-1082, 2009, doi: 10.1016/j.co