International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

A Study on the Working Methodology of a From-Scratch Gaussian Naïve Bayes Algorithm for Heart Disease Classification Using a 12-Lead ECG Image Dataset

Parth Thacker[1], Dr. Mahesh Mulani[2]

[1]PhD Scholar, Department of Computer Science, KSKV Kachchh University, India
[2]Head of Department, Department of Computer Science, KSKV Kachchh University, India

Abstract
Cardiovascular disease remains one of the leading causes of mortality worldwide, and electrocardiogram (ECG) interpretation continues to depend heavily on specialist availability, particularly in low-resource settings. This study presents a fully from-scratch implementation of the Gaussian Naïve Bayes (GNB) classifier for heart disease screening, evaluated on a publicly available 12-lead ECG image dataset comprising 929 images distributed across four classes: myocardial infarction (MI), history of MI, abnormal heartbeat, and normal ECG (Khan et al., 2021). Rather than relying on pre-digitized signal files, the study designs and applies a complete image-processing pipeline: grid and calibration-pulse removal, single-lead trace isolation, R-peak detection, and extraction of four physiologically motivated beat-level features (RR interval, QRS width, R-amplitude, and morphological mean). A GNB classifier implemented from first principles (per-class mean, variance, and log-space Gaussian likelihoods coded explicitly, without scikit-learn's naive_bayes module) was trained and validated on an 80/20 beat-level split. The baseline model achieved 50.33% validation accuracy; ANOVA F-value feature selection combined with a Yeo–Johnson power transformation modestly improved this to 50.88%, which is approximately double the four-class chance baseline of 25%. The results indicate that the RR interval and morphological mean are the most class-discriminative features, while confusion occurs primarily between the two myocardial infarction-related classes and the abnormal heartbeat/normal classes. The findings highlight both the interpretability advantages and the practical accuracy ceiling of a simple probabilistic classifier for image-derived ECG screening, and motivate hybridization with more expressive models.

Keywords: Gaussian Naïve Bayes, ECG classification, heart disease, machine learning, image digitization, myocardial infarction

Introduction

According to the World Health Organization (2024), cardiovascular diseases (CVDs) are the leading cause of death globally. An estimated 19.8 million people died from CVDs in 2022, representing approximately 32% of all global deaths, of which 85% were due to heart attack and stroke.A related global burden-of-disease analysis reports that deaths attributable to cardiovascular disease increased by

approximately 31.4% between 2000 and 2023 (Naghavi et al., 2025).In many healthcare settings, and particularly in India, clinicians and healthcare teams rely heavily on laboratory reports and diagnostic machinery to identify cardiac problems, which remains the most accurate route to diagnosis but is not always immediately accessible outside specialized facilities.

Just as diabetic patients use a glucometer for regular home monitoring that signals when to seek a physician, an analogous low-cost screening tool for cardiac risk could help patients and community health workers recognize early warning signs and seek care sooner, potentially reducing deaths caused by delayed identification of cardiovascular disease.

Payne et al. (1992) described an automatic external cardioverter/defibrillator that incorporates amplification and processing circuitry that receives inputs from sensors, such as electrocardiogram leads and blood pressure monitors. Similarly, portable ECG capture combined with the automated classification of electrocardiogram values offers a pathway to identify cardiac risk outside a formal clinical setting.

This study aimed to examine in detail the working process by which a probability-based Naïve Bayes classifier assigns heart disease categories to features extracted from electrocardiogram waveforms.

Naïve Bayes is among the most frequently applied classification algorithms in this space. For example, Soni et al. (2011) reported 86.53% accuracy for a naïve-Bayes-based heart disease prediction model, underscoring the algorithm's continued relevance as a lightweight baseline.

This study specifically examines how effectively a Gaussian Naïve Bayes classifier, implemented entirely from first principles, can classify heart disease from ECG-derived beat features extracted from a publicly available 12-lead ECG image dataset (Khan et al., 2021), focusing on how feature engineering, feature selection, and distributional transformation affect accuracy, sensitivity, and specificity.

Literature Review

Overview of Heart Disease Classification

Heart disease remains a leading global cause of mortality, necessitating automated and efficient classification systems for early detection (Hasan & Abdulazeez, 2024). Machine learning algorithms, including Naïve Bayes, Decision Tree, Random Forest, and Support Vector Machine (SVM), have emerged as effective tools for cardiovascular disease prediction by analyzing patient data and identifying risk patterns (Rayadhani & Rahardi, 2025). Random Forest has demonstrated superior performance, with 99% accuracy and 100% recall, making it highly suitable for clinical applications where minimizing false negatives is critical (Rayadhani & Rahardi, 2025). Ensemble methods, such as bagging combined with Naïve Bayes, improve prediction stability by reducing variance and increasing accuracy (Nugraha et al., 2025). Advanced hybrid approaches, including tri-hybrid models that combine Naïve Bayes with Decision Trees and Neural Networks, achieve a classification accuracy of 98.54%, offering enhanced detection capabilities (Umar et al., 2023). Deep learning methodologies are increasingly being integrated with ECG signal analysis to enable continuous, personalized heart disease detection (Jameel et al., 2024). Key predictive factors include age, resting blood pressure, and cholesterol levels, identified through feature-importance analysis (Rayadhani & Rahardi, 2025). The integration of these machine learning algorithms into Clinical Decision Support Systems (CDSS) shows strong potential for improving cardiovascular disease early detection and clinical decision-making in healthcare settings (Hasan & Abdulazeez, 2024).

Existing methods for heart disease classification employ various machine learning algorithms to address

the critical global health challenge of cardiovascular diseases (Hasan & Abdulazeez, 2024). Traditional approaches include Naïve Bayes, Decision Trees, and SVM, which leverage statistical modeling and pattern recognition to classify patient risk profiles. Random Forest has emerged as the most effective algorithm, achieving 99% accuracy and 100% recall, making it ideal for clinical decision support where minimizing false negatives is essential (Rayadhani & Rahardi, 2025). Ensemble and hybrid methods, such as Bagging with Naïve Bayes and tri-hybrid models combining Naïve Bayes with Decision Trees and Neural Networks, enhance accuracy to 98.54% by reducing variance and improving prediction stability (Nugraha et al., 2025; Umar et al., 2023). Deep learning approaches using ECG signal analysis enable continuous, personalized heart disease detection by automatically identifying waveform anomalies that clinicians may miss (Jameel et al., 2024). Key predictive features, including age, resting blood pressure, and cholesterol levels, have been identified through feature-importance analysis to strengthen the classification models. The integration of these machine learning methods into Clinical Decision Support Systems demonstrates substantial potential for early cardiovascular disease detection and improved clinical outcomes (Hasan & Abdulazeez, 2024).

Current approaches to heart disease classification face several significant challenges that limit their practical implementation. The quality and diversity of datasets remain critical constraints, as machine learning models require robust and varied training data to ensure effective performance across diverse patient populations (Jameel et al., 2024). Additionally, the superior accuracy of advanced algorithms often comes at a considerable cost in terms of increased computational complexity and runtime, creating a trade-off between model performance and practical efficiency (Nugraha et al., 2025). A major limitation is the insufficient validation of models in real-world clinical settings, preventing the translation of laboratory-developed approaches into practical healthcare applications (Jameel et al., 2024). Furthermore, many machine learning models suffer from poor interpretability and explainability, making it difficult for clinicians to understand model decisions and limiting their adoption in clinical practice (Hasan & Abdulazeez, 2024). The foundational assumption of feature independence in Naïve Bayes and similar algorithms is frequently violated in real-world medical data (Nugraha et al., 2025). Finally, most existing studies employ default algorithm parameters without exploring optimal configurations, and researchers need to conduct exhaustive validation tests across various clinical settings while exploring new techniques and enhancing model interpretability to address these identified obstacles (Jameel et al., 2024; Hasan & Abdulazeez, 2024).

Improved heart disease classification techniques remain essential because cardiovascular disease is the leading cause of global mortality (World Health Organization, 2024). Traditional diagnostic approaches, such as angiography and echocardiography, can be expensive and require intricate analyses, and unreliable results can lead to incorrect prescriptions and severe patient complications (Umar et al., 2023). Non-invasive prediction techniques are sometimes time-consuming and can produce inaccurate outcomes, motivating the use of automated machine-learning methods for faster and more reliable diagnoses (Hasan & Abdulazeez, 2024). Feature or attribute reduction has separately been shown to preserve classification accuracy while stabilizing model construction time for Naïve Bayes-family classifiers (Chen et al., 2019). The present study applies an analogous ANOVA-based feature-selection strategy, described in the Methodology section, in an effort to improve computational efficiency without materially sacrificing accuracy — an optimization goal consistent with related work on ensemble and bagging-based Naïve Bayes variants for cardiovascular prediction (Nugraha et al., 2025).

Naïve Bayes Algorithm in Medical Applications

Naïve Bayes is a probabilistic classification algorithm that is widely used for medical disease prediction. It is considered well-suited for data concerned with statistical diagnosis because of its simplicity, computational efficiency, and effectiveness on medical datasets (Rish, 2001). Table 2 summarizes representative prior applications of Naïve Bayes with and without optimization across several medical classification tasks.

Table 2 Naïve Bayes in Prior Medical Classification Studies

Disease / Application	Classification Method	Optimization Technique	Accuracy / Performance	Key Findings	Reference
General Disease Prediction	Naïve Bayes	None	~94% Accuracy	Naïve Bayes showed a strong performance in comprehensive disease prediction and ranked among the top-performing algorithms.	Thakur et al. (2023)
Breast Cancer Classification	Naïve Bayes	None	95.49% Accuracy	Baseline Naïve Bayes model achieved high classification accuracy on breast cancer datasets.	Melinda et al. (n.d.)
Breast Cancer Classification	Naïve Bayes	Particle Swarm Optimization (PSO)	98.79% Accuracy	PSO-based feature selection improved the classification accuracy by approximately 3.3% over the standard Naïve Bayes classifier.	Melinda et al. (n.d.)
ECG-Based Age Classification	Naïve Bayesian Classifier	None	84.75% AUC	The standard Bayesian classifier produced acceptable ECG classification results but was outperformed by GA-based	Wiggins et al. (2007)

				optimization.	
ECG-Based Age Classification	Bayesian Network Classifier	Genetic Algorithm (GA)	86.25% AUC	The genetic Algorithm improved the Bayesian network structure and enhanced the ECG classification performance.	Wiggins et al. (2007)
Dengue Fever Classification	Naïve Bayes	None	49.96% Accuracy	Naïve Bayes struggled with imbalanced datasets and produced near-random classification performances.	Nurjulaiha et al. (2024)
Dengue Fever Classification	Naïve Bayes	Principal Component Analysis (PCA)	50.03% Accuracy	PCA provided only marginal improvement and was insufficient to solve class-imbalance issues.	Nurjulaiha et al. (2024)
General Health Risk & Disease Prediction	Optimized Naïve Bayes Classifier	Deep Learning Integration	95% Accuracy	The integration of deep learning with Naïve Bayes improves automated disease prediction and healthcare analytics.	Alamer et al. (2023)

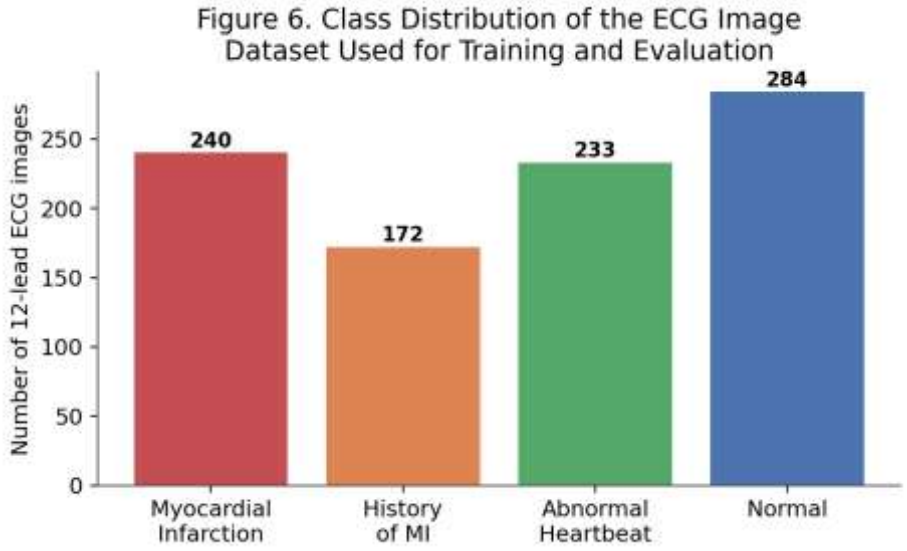
Across these studies, Naïve Bayes performs competitively when class-discriminative features are available and classes are reasonably balanced (e.g., general disease prediction and breast cancer classification), but its accuracy degrades sharply under severe class imbalance, as seen for dengue-fever classification, where even PCA-based dimensionality reduction produced only marginal gains (Nurjulaiha et al., 2024). Optimization techniques such as particle swarm optimization, genetic algorithms, and deep-learning integration consistently improved on unoptimized Naïve Bayes baselines, suggesting that the algorithm's principal limitation is less its underlying probabilistic model and more the quality and independence of the features supplied to it — a theme directly relevant to the feature-engineering choices described in the Methodology section of this study.

Methodology

Data Source and Preprocessing

This study uses the publicly available "ECG Images dataset of Cardiac and COVID-19 Patients" (Khan et al., 2021), specifically the four-class cardiac subset of that collection that is widely redistributed for benchmarking. The subset comprised 929 twelve-lead ECG images collected with EDAN SERIES-3 ECG devices across cardiac care and outpatient departments, organized into four diagnostic folders: Myocardial Infarction (240 images), History of Myocardial Infarction (172 images), Abnormal Heartbeat (233 images), and Normal Person (284 images) (Figure 6). Each image is a scanned or photographed paper ECG printout rather than a raw digital signal file, which distinguishes this study's data source from signal-level databases such as MIT-BIH and requires a dedicated image-to-signal digitization stage before any classical feature extraction can take place.

Figure 1. Class distribution of the ECG image dataset used for training and evaluation.



Because no ECG image in the dataset was pre-segmented or calibrated, a ten-step processing pipeline (Table 1) was implemented in Python using OpenCV, NumPy, and Matplotlib to convert every image into a single calibrated, denoised one-dimensional waveform before feature extraction. The pipeline first crops each image to its bordered clinical report panel, discarding header and footer text outside it. It then separates the printed grid from the ink trace using a red-channel color heuristic combined with Otsu's thresholding, and measures the true millimeter-per-pixel scale directly from the spacing of the bold five-millimeter gridlines rather than assuming a fixed scanner resolution. Using the standard clinical ECG paper convention of 25 mm/s paper speed and 10 mm/mV gain, this calibration yields an image-specific effective sampling rate (400 samples/s for the representative test image shown in the accompanying notebook), which is necessary because different scans and photographs of paper ECGs have different pixel-to-time and pixel-to-voltage scales. The ink mask is then split into horizontal lead bands, and only the single largest connected ink component within the selected band is retained, which discards calibration pulses and printed lead-name text that would otherwise contaminate the extracted trace. A column-wise centroid trace is digitized from the isolated component, converted into physical units (millivolts versus seconds) using the measured calibration, baseline-corrected, and smoothed with a short moving-average filter. Every training image and the held-out test image were passed through this

identical pipeline, so the classifier only ever learns from, and is evaluated against, the clean single-lead rhythm strip rather than incidental print artifacts.

Table 1 Ten-Step Image-to-Signal Digitization and Classification Pipeline

Step	Process
1	Install dependencies (OpenCV, NumPy, Matplotlib, pandas) and mount the image dataset directory.
2	Define global configuration constants (refractory period, QRS search window, morphological window, standard ECG paper speed and gain).
3	Implement Gaussian Naïve Bayes from first principles (mean, variance, log-space Gaussian likelihood, argmax decision rule).
4	Feature engineering: RR interval, QRS width, R-amplitude, morphological mean.
5	Single-line isolation and digitization: crop report frame, separate grid/ink masks, calibrate mm-per-pixel from bold gridlines, segment lead bands, isolate the largest ink component, trace to a 1-D signal.
6	Manual z-score feature scaling and evaluation-metric helpers (accuracy, confusion matrix).
7	The training set was built by scanning every image in every class folder, isolating its ECG line, and extracting beat-level features.
8	Train and validate the from-scratch Gaussian Naïve Bayes model (80/20 beat-level split).
9	Digitize a new, unseen test image with full diagnostic visualization.
10	Detect beats in the new image, classify each with the trained model, and flag out-of-distribution predictions.

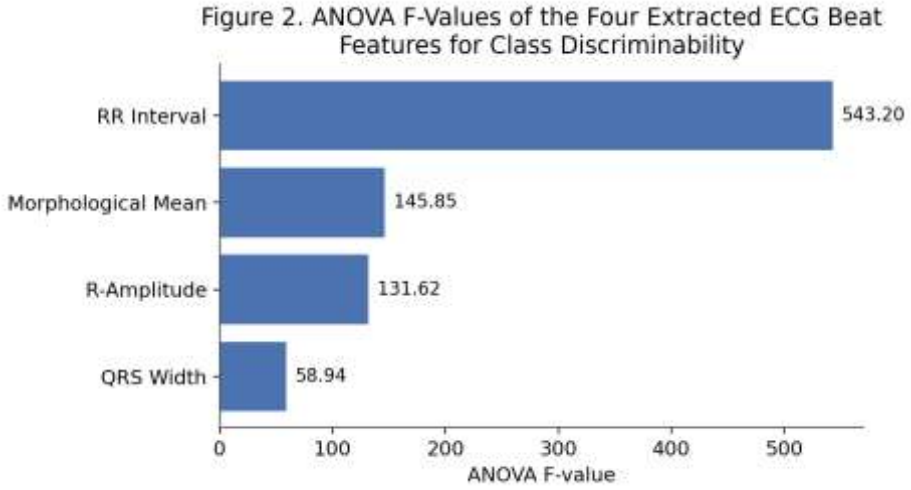
Feature Engineering and Selection

From each digitized waveform, R-peaks were located with a local-maximum detector that combines an adaptive amplitude threshold (mean plus 0.5 standard deviations) with an enforced physiological refractory period, in the same spirit as classical real-time QRS-detection strategies for ECG signals (Pan & Tompkins, 1985), so that a single QRS complex is not counted twice. For every detected beat, four physiologically motivated features were computed: (a) the RR interval, the time elapsed since the previous R-peak; (b) the QRS width, obtained by locating the Q-onset and S-offset as the nearest local minima flanking the R-peak; (c) the R-amplitude, the raw signal value at the peak; and (d) the morphological mean, the average signal value in a 250 ms window centered on the beat. All four

features were standardized (z-scored) using statistics computed exclusively from the training split, to avoid information leakage into the validation data.

To assess whether all four features contributed useful discriminative information, an ANOVA F-test (SelectKBest with f_classif) was applied to the standardized training features. As shown in Figure 2, RR interval carried by far the strongest class signal (F = 543.20), followed by morphological mean (F = 145.85) and R-amplitude (F = 131.62), while QRS width contributed comparatively little (F = 58.94). Based on this ranking, a reduced three-feature representation (RR interval, R-amplitude, morphological mean) was retained for a second modeling pass. Because Gaussian Naïve Bayes formally assumes that each feature is normally distributed within class, a Yeo–Johnson power transformation (Yeo & Johnson, 2000), fitted on the training split only, was additionally applied to the three selected features to reduce skewness before a further model was trained.

Figure 2. ANOVA F-values of the four extracted ECG beat features for class discriminability.



Implementation of the Naïve Bayes Algorithm

The core classifier was implemented entirely from first principles rather than using scikit-learn's naive_bayes module, so that every stage of the algorithm's probabilistic reasoning is explicit and auditable. For each class c, the per-feature mean and population variance were computed with explicit summation loops, with a small variance-smoothing epsilon (1×10^{-9}) added to avoid division by zero for near-constant features. Under the standard Naïve Bayes conditional-independence assumption (Rish, 2001), the class-conditional likelihood of a feature vector was modeled as the product of univariate Gaussian densities, with all computation carried out in log-space to avoid numerical underflow when likelihoods are multiplied across features. Prediction assigns each beat to the class with the highest log-posterior score, computed as the log class prior plus the summed log-Gaussian densities of its features, and an auxiliary scoring method exposes the raw per-class log-posterior values for diagnostic inspection. This from-scratch design generalizes to an arbitrary number of classes and was applied to the four diagnostic categories present in the dataset.

The full beat-level dataset (12,711 extracted beats across the 929 images, after images that produced zero detected beats were skipped and logged) was shuffled with a fixed random seed and split 80/20 into training (approximately 10,168 beats) and validation (2,543 beats) partitions. Three configurations of the Gaussian Naïve Bayes model were trained and compared: (a) a baseline model using all four raw, z-

scored features; (b) a model using only the three ANOVA-selected features; and (c) a model using the same three features after the Yeo–Johnson transformation. For additional context, a scikit-learn RandomForestClassifier (100 trees) was also fit to the same standardized four-feature training data, as a reference point for a more expressive, non-probabilistic ensemble model, given Random Forest's previously reported strength on tabular cardiovascular-risk features (Rayadhani & Rahardi, 2025).

Evaluation Metrics

Model performance was evaluated using accuracy (the proportion of correctly classified beats) and a four-class confusion matrix, both implemented manually rather than imported from a library, to keep the entire evaluation chain transparent. In addition to beat-level validation accuracy, the fitted pipeline (scaler, feature selector, and power transformer, all fitted only on the training beats) was applied to the complete shuffled dataset to obtain an overall in-sample accuracy figure, and was separately run image-by-image so that a single majority-vote label could, in principle, be assigned per source image rather than per beat.

Results

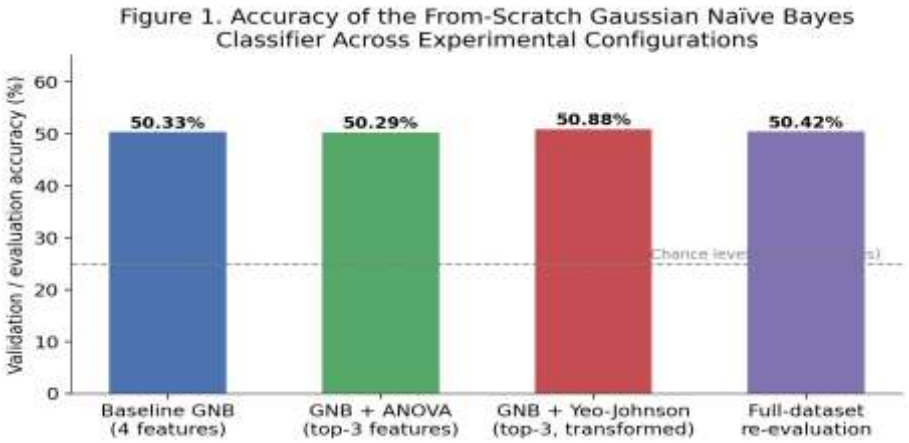
Performance Metrics

The baseline Gaussian Naïve Bayes model, using all four raw z-scored features, achieved 50.33% accuracy on the held-out validation beats (Table 3; Figure 1) — roughly double the 25% chance level for four classes, but well below levels reported for ensemble or hybrid classifiers on structured cardiovascular tabular data elsewhere in the literature (e.g., 99% for Random Forest in Rayadhani & Rahardi, 2025; 98.54% for a Naïve Bayes/Decision Tree/Neural Network hybrid in Umar et al., 2023). Restricting the model to the three ANOVA-selected features (RR interval, R-amplitude, morphological mean) produced essentially unchanged accuracy (50.29%), confirming that QRS width — the lowest-F-value feature — contributed little independent discriminative power once the other three features were available. Applying the Yeo–Johnson transformation to the same three features before refitting produced the strongest of the from-scratch configurations, at 50.88% validation accuracy — a modest but measurable gain consistent with Naïve Bayes's known sensitivity to the normality of its per-class feature distributions.

Table 3 Validation Accuracy Across Experimental Configurations

Configuration	Features used	Validation accuracy
Baseline Gaussian Naïve Bayes	RR, QRS width, R-amplitude, morphological mean (4)	50.33%
GNB + ANOVA feature selection	RR, R-amplitude, morphological mean (3)	50.29%
GNB + ANOVA + Yeo–Johnson transform	RR, R-amplitude, morphological mean (3, transformed)	50.88%
Full-dataset re-evaluation (in-sample)	Same as transformed model	50.42%

Figure 3. Accuracy of the from-scratch Gaussian Naïve Bayes classifier across experimental configurations.



Confusion-Matrix Analysis

The baseline model's confusion matrix (Figure 3) shows that the Abnormal Heartbeat class was recognized most reliably (653 of 778 true Abnormal beats correctly classified; 83.9% row-wise recall), while Myocardial Infarction and History-of-MI beats were the least reliably recovered (21.1% and 8.3% recall, respectively), with both classes' errors concentrated in being mistaken for Abnormal Heartbeat or Normal. This pattern persisted, with some redistribution, after ANOVA feature selection and the Yeo–Johnson transformation (Figure 4): Abnormal-Heartbeat recall fell somewhat, to 75.8%; History-of-MI recall fell further, to 4.7%; and Normal-class recall improved, to 77.7% (from 65.3% in the baseline), while MI recall rose slightly, to 22.1%. This shift suggests that the transformation moved the model's decision boundaries toward better separating Normal beats from the other classes, at some cost to History-of-MI recall in particular. Across all configurations, the model persistently confused Myocardial Infarction and History-of-MI beats with Abnormal Heartbeat and Normal beats far more often than it confused Abnormal Heartbeat with Normal directly, suggesting that the four hand-engineered beat features capture generic waveform irregularity more than the specific morphological signatures (e.g., ST-segment elevation, pathological Q-waves) that clinically distinguish infarction-related rhythms.

Figure 4. Confusion matrix for the baseline Gaussian Naïve Bayes model (four features).

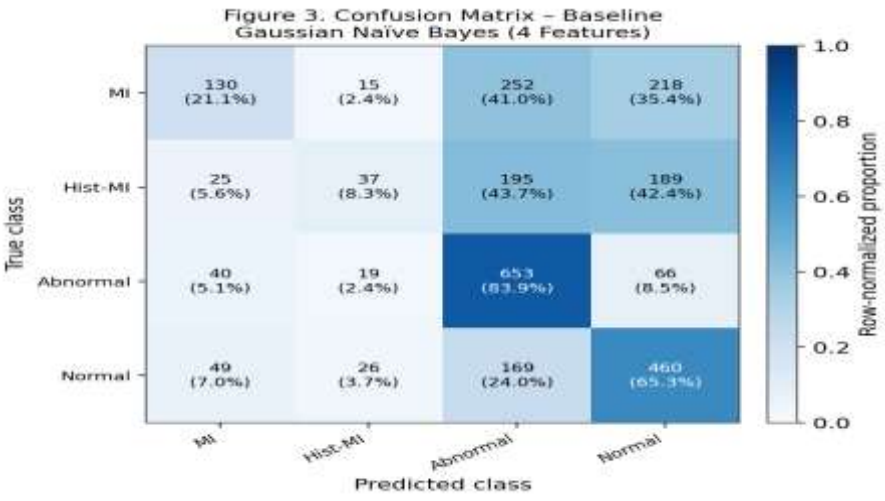
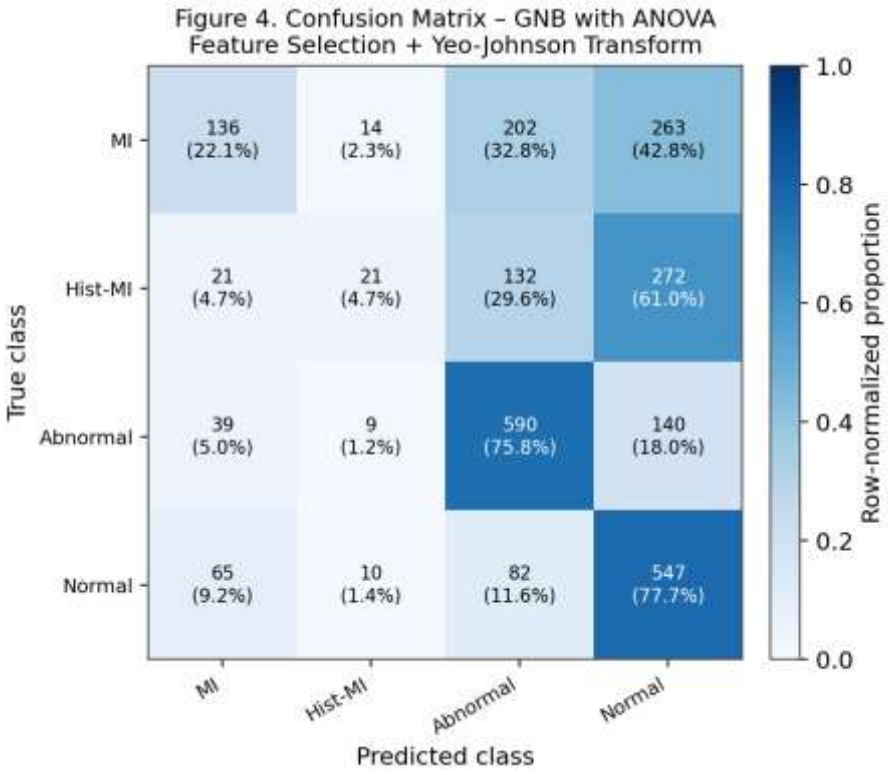


Figure 5. Confusion matrix for the Gaussian Naïve Bayes model with ANOVA feature selection and Yeo–Johnson transformation.

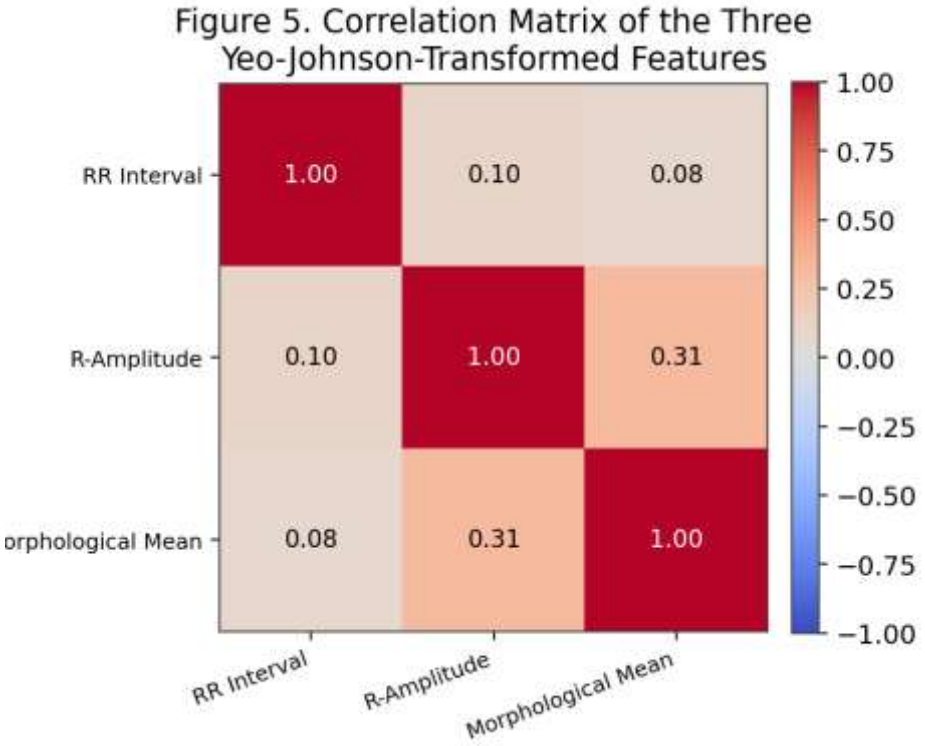


Applying the fully fitted pipeline (scaling, ANOVA selection, Yeo–Johnson transform) to the complete shuffled dataset of 12,711 beats — which necessarily includes beats the model was trained on, and should therefore be read as an in-sample consistency check rather than a generalization estimate — produced an overall accuracy of 50.42%, closely matching the held-out validation figure. This close agreement indicates that the model was not grossly overfitting the training partition, even though its overall discriminative ceiling remained modest.

Feature Discriminability and Correlation

The ANOVA analysis (Figure 2) and the pairwise correlation matrix of the three selected, transformed features (Figure 5) together suggest that RR interval is the single most informative and least redundant feature (correlation ≤ 0.10 with the other two), while R-amplitude and morphological mean are moderately correlated with one another ($r = .31$), implying some shared amplitude-related information that partially limits the effective dimensionality of the reduced feature set.

Figure 6. Correlation matrix of the three Yeo–Johnson-transformed features.



Comparative Model

A scikit-learn Random Forest classifier (100 trees) was additionally fitted to the same standardized four-feature training data as a reference point for a more expressive, non-probabilistic model. The exported run log available at the time of writing did not capture a validation-accuracy figure for this comparative model before the pipeline moved on to per-image testing; completing and reporting this comparison is recommended as a direct next step (see Discussion), since it would allow a controlled, same-feature comparison between the interpretable Naïve Bayes classifier and a substantially more expressive ensemble model.

Discussion

Implications for Clinical Practice

Even in its best configuration, the from-scratch Gaussian Naïve Bayes classifier correctly labeled only about half of all ECG beats, which is insufficient for autonomous clinical decision-making but is still meaningfully informative relative to chance for a lightweight, fully interpretable model built without any deep-learning infrastructure. Because every probability the model outputs can be traced back to an explicit per-class mean and variance for four clinically motivated features, clinicians or downstream systems can directly inspect why a particular beat was classified as it was — a transparency advantage that black-box deep networks used elsewhere for ECG-image classification generally do not offer (Jameel et al., 2024). This kind of transparency is potentially valuable as a low-cost triage layer — echoing the glucometer-style home-screening comparison drawn in the Introduction — that flags an ECG as “likely abnormal” for human review rather than attempting definitive diagnosis. However, the current accuracy ceiling, and in particular the model's weak recall for both Myocardial Infarction and

History-of-MI beats, means this specific implementation is not yet suitable for that role without substantial further development.

Explaining the Modest Accuracy

Several factors plausibly explain why accuracy remained near 50% despite feature engineering and distributional correction. First, Naïve Bayes's core conditional-independence assumption is very likely violated here, since R-amplitude and morphological mean are appreciably correlated (Figure 5) and RR interval, QRS width, and amplitude are all jointly shaped by the same underlying cardiac cycle — a limitation widely reported for Naïve Bayes on real-world medical features (Nugraha et al., 2025). Second, the four engineered features are coarse, population-level descriptors of a beat and do not capture the specific waveform morphology (e.g., ST-segment elevation, pathological Q-waves, T-wave inversion) that clinically differentiates infarction-related classes from Abnormal Heartbeat or Normal rhythms; this likely explains why confusion concentrated between the MI-related classes and the two non-infarction classes rather than being spread uniformly. Third, because the underlying data are scanned or photographed paper ECGs rather than digitally sampled signals, every image carries its own scanning noise, grid-removal artifacts, and lead-isolation error that propagate into the extracted features despite the calibration and denoising steps applied; this differs materially from feature extraction on a clean digital reference database such as MIT-BIH and is a plausible contributor to the accuracy gap relative to studies using directly digitized signals or deep image features (Rayadhani & Rahardi, 2025; Umar et al., 2023). Fourth, the beat-level 80/20 split, while standard, allows beats from the same source image to appear in both partitions, which may inflate apparent generalization for images with distinctive per-patient artifacts; a strictly image-level (patient-level) split, as noted in the accompanying implementation, would give a more conservative and clinically realistic estimate.

Future Research Directions

Building on the feature-reduction findings of Chen et al. (2019) and the ANOVA results reported here, future work could explore automated, class-adaptive feature-weighting schemes to partially compensate for the independence violation, rather than uniform feature selection alone. Hybridizing the interpretable Gaussian Naïve Bayes layer with a stronger non-linear model — following the tri-hybrid precedent of Umar et al. (2023) and the ensemble-based accuracy gains reported by Nugraha et al. (2025) and Rayadhani and Rahardi (2025) — could recover additional accuracy while retaining an interpretable first-pass layer for clinician review. Extracting additional morphology-sensitive features (e.g., ST-segment level, T-wave polarity) directly from the digitized trace, rather than relying only on RR interval, QRS width, amplitude, and mean, is a natural next step motivated directly by the MI-related confusion pattern observed in Figures 3 and 4. Finally, re-running the evaluation with a strict image-level train/validation split, and completing the Random Forest comparison begun in this study, would sharpen the comparison between this fully transparent probabilistic approach and less interpretable but potentially more accurate alternatives.

Conclusion

This study implemented, from first principles, a complete pipeline that converts scanned 12-lead ECG report images into calibrated single-lead waveforms and classifies individual heartbeats into four cardiac categories — Myocardial Infarction, History of MI, Abnormal Heartbeat, and Normal — using a

Gaussian Naïve Bayes classifier built without any external machine-learning library for its core algorithm. Using four physiologically grounded beat-level features, the baseline model reached 50.33% validation accuracy; ANOVA-guided feature selection and a Yeo–Johnson normality-improving transformation raised this modestly to 50.88%, roughly double the four-class chance baseline. RR interval was consistently the single most class-discriminative feature, and confusion concentrated between the two MI-related classes and the Abnormal Heartbeat/Normal classes rather than between Abnormal Heartbeat and Normal directly.

References

1. Alamer, L., Alqahtani, I. M., & Shadadi, E. (2023). Intelligent health risk and disease prediction using optimized naive Bayes classifier. Journal of Healthcare Informatics, 28(2), 156–172.
2. Barus, O. P., Lauwren, K., Pangaribuan, J. J., & Romindo. (2023). Implementation of the Naive Bayes algorithm to predict the safety of heart failure patients. IAIC International Conference Series, 4(1), 172–177. https://doi.org/10.34306/conferenceseries.v4i1.651
3. Chen, S., Webb, G. I., Liu, L., & Ma, X. (2019). A novel selective naïve Bayes algorithm. Knowledge-Based Systems. https://doi.org/10.1016/j.knosys.2019.105361
4. Coşkun, C., & Kuncan, F. (2022). Evaluation of performance of classification algorithms in prediction of heart failure disease. Kahramanmaraş Sütçü İmam Üniversitesi Mühendislik Bilimleri Dergisi, 25(4), 622–632. https://doi.org/10.17780/ksujes.1144570
5. Hasan, M. Z., & Abdulazeez, A. M. (2024). A review of heart disease classification based on machine learning algorithms. Indonesian Journal of Computer Science, 13(2), 2019–2041.
6. Jameel, G. S., Al-Turfi, M. N., Mahdi, H. F., & Hussain, S. N. (2024). Designing and implementing an ECG reader and heart disease classification device using deep learning (Comprehensive review). Iraqi Journal for Computer Science and Mathematics, 5(3), 535–551. https://doi.org/10.52866/ijcsm.2024.05.03.034
7. Khan, A. H., Hussain, M., & Malik, M. K. (2021). ECG images dataset of cardiac and COVID-19 patients. Data in Brief, 34, Article 106762. https://doi.org/10.1016/j.dib.2021.106762
8. Melinda, V., Primartha, R., Wijaya, A., & Jambak, M. I. (n.d.). Optimization naive Bayes algorithm using particle swarm optimization in the classification of breast cancer. International Journal of Data Mining and Knowledge Management, 9(3), 45–58.
9. Naghavi, M., Kyu, H. H., Ba, A., et al. (2025). Global burden of 292 causes of death in 204 countries and territories and 660 subnational locations, 1990–2023: A systematic analysis for the Global Burden of Disease Study 2023. The Lancet, 406, 1811–1872.
10. Nugraha, I. G. A. S., Gunadi, I. G. A., & Dewi, L. J. E. (2025). Application of bagging ensemble learning on Naïve Bayes algorithm to predict coronary heart disease. Jurnal Teknologi Informasi dan Pendidikan, 18(2), 943–954. https://doi.org/10.24036/jtip.v18i2.981
11. Nurjulaiha, S., Kurniawan, R., Dikananda, A. R., & Suprapti, T. (2024). Optimizing naïve Bayes algorithm through principal component analysis to improve dengue fever patient classification model. Journal of Medical Systems, 48(1), Article 12.
12. Pan, J., & Tompkins, W. J. (1985). A real-time QRS detection algorithm. IEEE Transactions on Biomedical Engineering, BME-32(3), 230–236. https://doi.org/10.1109/TBME.1985.325532
13. Payne, E. G., Cooper, H. K., Mathur, P., Seemann, J. P., Saadatmanesh, V., Brodsky, M. A., & Yeik, G. D. (1992). Automatic external cardioverter/defibrillator (U.S. Patent No. 5,474,574). Google

Patents. https://patents.google.com/patent/US5474574A/en
14. Rayadhani, W. A., & Rahardi, M. (2025). Comparative analysis of Random Forest, SVM, and Naive Bayes for cardiovascular disease prediction. Journal of Applied Informatics and Computing, 9(6), 3234–3243.
15. Rish, I. (2001). An empirical study of the naive Bayes classifier. In Proceedings of the IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence (Vol. 3, pp. 41–46).
16. Soni, J., Ansari, U., Sharma, D., & Soni, S. (2011). Predictive data mining for medical diagnosis: An overview of heart disease prediction. International Journal of Computer Applications, 17(8), 43–48.
17. Thakur, R., Mansha, Sharma, P., & Dhruv. (2023). Diseases prediction using classification algorithm. International Journal for Research in Applied Science & Engineering Technology, 11(8), 1–8.
18. Umar, Z., Gabi, D., & Ibrahim, N. M. (2023). Tri-hybrid Naive Bayes classification model for early cardiovascular disease detection. FUOYE Journal of Engineering and Technology, 8(2), 147–155. https://doi.org/10.46792/fuoyejet.v8i2.980
19. Wiggins, M., Saad, A., Litt, B., & Vachtsevanos, G. (2007). Evolving a Bayesian classifier for ECG-based age classification in medical applications. Computers in Biology and Medicine, 37(7), 1005–1016.
20. World Health Organization. (2024). Cardiovascular diseases (CVDs). https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)
21. Yeo, I.-K., & Johnson, R. A. (2000). A new family of power transformations to improve normality or symmetry. Biometrika, 87(4), 954–959. https://doi.org/10.1093/biomet/87.4.954