

Benchmarking Machine Learning and Ensemble Techniques for Accurate Fetal Health Classification Using Cardiotocogram Data.

Dr. Pankaj Richhariya¹, Jeya Fatima^{2,*}

¹Head of Department, Computer Science and Engineering, BITS Bhopal, Bhopal, India

²Department of Computer Science and Engineering, BITS Bhopal, Bhopal, India

*Corresponding Author E-mail: zeyafatima3265@gmail.com

Abstract

Cardiotocography (CTG) is a widely used non-invasive technique for monitoring fetal heart rate and uterine contractions during pregnancy, providing critical information to detect fetal distress and prevent adverse outcomes. This study benchmarks multiple machine learning and ensemble techniques for automated, accurate classification of fetal health status using CTG-derived features from the publicly available UCI Cardiotocography dataset, comprising 2126 records labeled as Normal, Suspect, or Pathological. Following preprocessing with feature scaling and an 80:20 stratified train-test split, five base classifiers were trained and evaluated: Logistic Regression, Support Vector Machine (SVM), Random Forest, K-Nearest Neighbors (KNN), and XGBoost. To further improve predictive performance, two ensemble strategies combining XGBoost and Random Forest were implemented: a soft-voting Blending classifier and a Stacking classifier with Logistic Regression as the meta-learner, validated using 5-fold cross-validation. Model performance was assessed using accuracy, precision, recall, F1-score, confusion matrices, ROC-AUC, and the Matthews Correlation Coefficient (MCC) to account for class imbalance. Results show XGBoost achieved the best standalone performance, with 96.00% accuracy and an MCC of 0.8893, slightly outperforming both the Blending and Stacking ensembles (95.77% accuracy, MCC of 0.8839 and 0.8829 respectively), while Random Forest followed closely at 95.00% accuracy. Logistic Regression yielded the weakest baseline at 88.00% accuracy. Although ensemble methods did not surpass standalone XGBoost in this experiment, Stacking demonstrated stable generalization with a mean cross-validation accuracy of 94.29%. These findings indicate that gradient boosting effectively captures the complex, imbalanced, multi-class structure of CTG data and holds strong potential for integration into clinical decision-support systems for early, reliable detection of fetal distress.

Keywords: *Fetal Health Classification, Cardiotocography, Machine Learning, Ensemble Learning, XGBoost, Class Imbalance, Matthews Correlation Coefficient.*

1. Introduction

Fetal health monitoring during pregnancy and labor is one of the central tasks of modern obstetric care, aimed at the early identification of complications that can lead to adverse maternal and neonatal outcomes. According to the World Health Organization, hundreds of thousands of stillbirths and neonatal deaths occur globally each year, a substantial proportion of which are potentially preventable with timely detection of fetal distress [1]. Cardiotocography (CTG) remains the standard non-invasive method used by clinicians to continuously record fetal heart rate (FHR) and uterine contractions (UC), enabling assessment of fetal well-being throughout antenatal monitoring and labor.

Despite its widespread clinical use, manual interpretation of CTG traces is inherently subjective. Studies have long documented substantial inter- and intra-observer variability among obstetricians reading the same trace, which can lead either to delayed intervention in genuinely distressed fetuses or to unnecessary intervention — including avoidable operative deliveries — in healthy ones. This variability motivates the search for objective, reproducible, and automated decision-support tools that can complement clinical judgement rather than replace it.

The application of machine learning (ML) to quantitative, expert-derived CTG features offers a promising path toward such objective and automated fetal health classification. Unlike raw waveform interpretation, structured numerical features summarizing FHR baseline, variability, accelerations, and decelerations can be fed directly into supervised learning models, enabling systematic benchmarking of algorithmic approaches on a common, controlled experimental protocol. However, CTG datasets are typically class-imbalanced, with the clinically critical Pathological class being far rarer than the Normal class. This imbalance complicates model training and demands evaluation metrics that go beyond raw accuracy — a concern this study addresses directly through the systematic use of the Matthews Correlation Coefficient (MCC) [9], [16].

This paper presents a comprehensive, mathematically grounded benchmarking study of five widely used machine learning classifiers — Logistic Regression, Support Vector Machine (SVM), Random Forest, K-Nearest Neighbors (KNN), and XGBoost — alongside two ensemble learning strategies (Blending and Stacking) for classifying fetal health into Normal, Suspect, and Pathological categories using the publicly available UCI Cardiotocography dataset. The principal contributions of this work are:

- A controlled, single-pipeline comparison of five classical and modern ML classifiers under identical preprocessing, splitting, and evaluation conditions, avoiding the

confounds of aggregating results across heterogeneous studies.

- Construction and evaluation of two complementary ensemble strategies (soft-voting Blending and Stacking with a Logistic Regression meta-learner) built on the two strongest base learners.
- A class-imbalance-aware evaluation protocol that reports MCC and per-class F1-score on the clinically critical Pathological class, rather than relying on accuracy alone.
- Full mathematical formulations for every algorithm and metric used, together with confusion-matrix-level analysis of model errors.

The remainder of this paper is organized as follows: Section 2 reviews related work; Section 3 describes the dataset and presents exploratory data analysis; Section 4 details the methodology, including full mathematical formulations of each algorithm; Section 5 presents results and discussion with visualizations; Section 6 discusses limitations; and Section 7 concludes with directions for future work.

Beyond its algorithmic contributions, this work is motivated by a practical clinical concern: in intrapartum monitoring, a classifier's most costly error is not a false alarm on a healthy fetus but a missed Pathological case, since the latter directly delays a potentially life-saving intervention. Consequently, this paper deliberately reports per-class F1-score on the Pathological category and inspects raw confusion matrices in addition to the aggregate accuracy and MCC figures typically emphasized in benchmarking studies, so that the clinical implications of each model's error profile are made explicit rather than hidden inside a single summary number.

2. Related Work

Prior studies on CTG-based fetal health classification have explored a wide range of machine learning approaches, from traditional linear classifiers such as Logistic Regression [11] to kernel-based methods such as Support Vector Machines, and from instance-based learners such as K-Nearest Neighbors [3] to tree-ensemble methods such as Random Forest and gradient boosting variants. Cortes and Vapnik's foundational work on support-vector networks established the margin-maximization principle that underlies SVM classification [2], while Breiman's random forest formulation demonstrated that aggregating decorrelated decision trees, grown on bootstrap samples with random feature subsets, substantially reduces variance relative to a single decision tree [4], building on the earlier CART framework of Breiman et al. [5]. More recently, gradient boosting frameworks such as XGBoost, introduced by Chen and Guestrin, have become a dominant approach for structured tabular data due to their regularized, additive tree-building procedure and strong empirical performance across many domains, including clinical prediction tasks [6].

A recurring theme in this literature is the class imbalance inherent in CTG datasets, where Pathological cases are considerably rarer than Normal cases; on the UCI Cardiotocography dataset used here, Normal, Suspect, and Pathological records occur in an approximate 78:14:8 ratio [17]. Because accuracy alone can be misleading under

imbalance, researchers increasingly report the Matthews Correlation Coefficient, which Matthews originally proposed for evaluating binary classifiers [9] and which has since been extended to multi-class settings as a single, balanced summary statistic that accounts for all four cells of the confusion matrix simultaneously. Chicco and Jurman further showed empirically that MCC is markedly more informative than accuracy or F1-score precisely in the imbalanced regime typical of biomedical classification tasks [16]. Complementary strategies for handling class imbalance at the data level, such as the Synthetic Minority Over-sampling Technique (SMOTE) of Chawla et al. [15], are widely used in related biomedical classification literature, though this study instead relies on metric-level correction via MCC, consistent with the original ICCSAM protocol.

Ensemble approaches combining multiple base learners — whether through simple averaging (Blending, i.e., soft voting) or through a trained meta-learner (Stacking, as formalized by Wolpert [7]) — have also been shown in the broader ensemble-learning literature (Dietterich [8]) to reduce variance and improve robustness relative to any single constituent model, motivating their inclusion in this study alongside standalone classifiers. Receiver-operating-characteristic analysis, following Fawcett [10], is used as a complementary discrimination measure alongside the confusion-matrix-derived statistics.

Closely related to the present work, Piri et al. proposed an ensemble of hyper-parameter-tuned classifiers for fetal health prediction on the same UCI Cardiotocography dataset, reporting substantial gains from tuning and ensembling relative to default-configuration base learners [17]. That work, like much of the surrounding literature, primarily emphasizes aggregate accuracy; the present study differs in its explicit, side-by-side reporting of MCC and per-class Pathological F1-score alongside accuracy, and in its use of a single shared preprocessing and evaluation pipeline across every classifier and ensemble so that differences in reported performance can be attributed to the algorithms themselves rather than to protocol differences.

This study builds on that foundation by systematically benchmarking five individual classifiers and two ensemble techniques (Blending and Stacking) on a common experimental pipeline — implemented with the scikit-learn library [14] — with identical preprocessing, an 80:20 stratified train-test split, and a shared set of evaluation metrics, enabling a direct, controlled comparison rather than aggregating results across heterogeneous studies with differing protocols.

It is worth noting that several prior CTG studies report accuracies in a similar 90–96% range using Random Forest or boosted-tree variants, but differ substantially in preprocessing choices (e.g., whether the 10-class morphologic labels or the 3-class fetal-state labels are used), in the fraction of data held out for testing, and in whether class-imbalance-aware metrics such as MCC are reported at all. This heterogeneity makes cross-paper comparison unreliable, which is precisely the gap the present controlled benchmark is designed to close: every classifier and ensemble reported here is trained and evaluated

inside the same pipeline, on the same 1700/426 split, with the same random seed and the same metric suite.

3. Dataset Description

The dataset used in this study is the publicly available Cardiotocography (CTG) dataset, distributed via the UCI Machine Learning Repository [13] and commonly redistributed on Kaggle as “fetal_health.csv,” consisting of 2126 fetal cardiotocogram recordings. Each record is described by 21 quantitative features derived from FHR and uterine contraction signals, summarized in Table 1, and is labeled by expert obstetricians into one of three fetal health classes: Normal, Suspect, or Pathological.

The dataset was checked for missing values prior to modelling; none were found across the 21 feature columns, so no imputation step was required. The dataset exhibits notable class imbalance, with Normal cases substantially outnumbering Suspect and Pathological cases (Fig. 2): 1655 records (77.8%) are labeled Normal, 295 (13.9%) Suspect, and 176 (8.3%) Pathological. This imbalance was explicitly accounted for in the choice of evaluation metrics (Section 4.9) and motivates reporting per-class performance on the minority Pathological class in addition to aggregate scores.

Each CTG recording is automatically summarized into 21 numerical descriptors spanning four broad categories (Table 1). The Baseline & Movement group captures the average fetal

heart rate and the frequency of accelerations, fetal movements, and uterine contractions per unit time. The Decelerations group counts light, severe, and prolonged decelerations, which are transient drops in FHR associated with cord compression or placental insufficiency. The Variability group summarizes short- and long-term fluctuation in FHR — clinically among the most informative signals, since a healthy autonomic nervous system produces continuous beat-to-beat variability, and its suppression is an early marker of fetal hypoxia. Finally, the Histogram Statistics group describes the overall shape of the FHR distribution across the recording (width, central tendency, skewness-related tendency, and peak/zero counts), effectively encoding a compressed view of the full FHR trace without requiring the raw waveform itself.

Because the twenty-one features differ by several orders of magnitude in scale — for instance, baseline heart rate is measured in beats per minute (commonly 110–160) while acceleration and deceleration counts are typically small integers or proportions close to zero — distance-based and gradient-based learners would otherwise be dominated by the highest-magnitude features. This motivates the standardization step formalized in Eq. (1) and applied uniformly across all seven models for consistency, even though tree-based methods such as Random Forest and XGBoost are, in principle, invariant to monotonic feature rescaling.

Category	Representative Features
Baseline & Movement	baseline value, accelerations, fetal_movement, uterine_contractions
Decelerations	light_decelerations, severe_decelerations, prolonged_decelerations
Variability	abnormal_short_term_variability, mean_value_of_short_term_variability, percentage_of_time_with_abnormal_long_term_variability, mean_value_of_long_term_variability
Histogram Statistics	histogram_width, histogram_min, histogram_max, histogram_mode, histogram_mean, histogram_median, histogram_variance, histogram_tendency, number_of_peaks, number_of_zeroes
Target	fetal_health $\in$ {1: Normal, 2: Suspect, 3: Pathological}

Table 1. Summary of CTG feature categories used as model input.

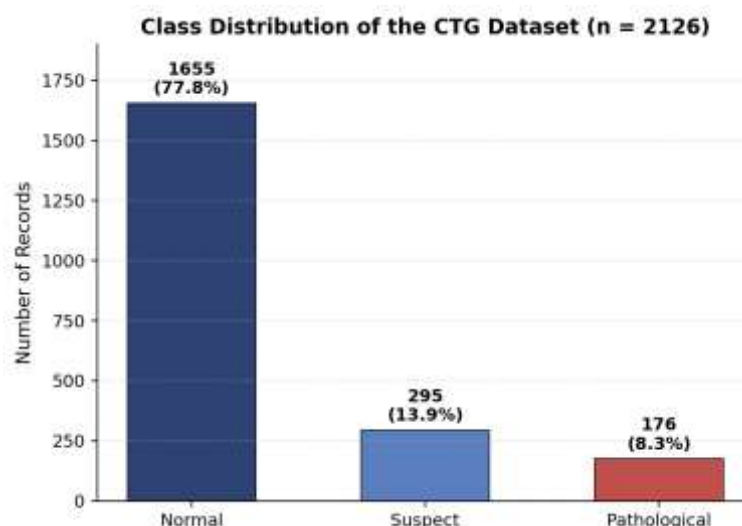


Fig. 2. Class distribution of the CTG dataset (n = 2126), showing pronounced imbalance toward the Normal class.

Exploratory data analysis (EDA) additionally included univariate distribution plots (histograms and box plots) for each of the 21 features, bivariate pairplots segmented by class label, and a Pearson correlation heatmap to identify multicollinearity among features. This analysis informed the feature-scaling decisions described in Section 4.1 and provided qualitative insight into class separability prior to model training. Consistent with prior CTG literature, features related to deceleration counts and short/long-term FHR variability — notably prolonged decelerations, abnormal short term variability, and percentage of time with abnormal long term variability — showed the strongest association with the Pathological class, in line with the clinical understanding that reduced heart-rate variability and prolonged decelerations are hallmark indicators of fetal compromise.

4. Methodology

Fig. 1 summarizes the end-to-end methodology, from raw CTG data through preprocessing, base model training, ensemble construction, and final model comparison and selection.

4.1 Data Preprocessing

The feature matrix $X \in \mathbb{R}^{2126 \times 21}$ and target vector y were constructed from the 21 input features and the three-class label. Features were standardized using z-score normalization (StandardScaler) so that each feature has zero mean and unit variance, which is particularly important for distance-based and gradient-based models such as KNN, SVM, and Logistic Regression:

$$x' = (x - \mu) / \sigma \quad (1)$$

where μ and σ are the mean and standard deviation of a feature computed on the training set. The data was then partitioned into training (80%, $n = 1700$) and testing (20%, $n = 426$) subsets using stratified random sampling to preserve class proportions (random_state = 42).

4.2 Exploratory Data Analysis

As described in Section 3, exploratory analysis combined univariate, bivariate, and correlation-based views of the data. This step is treated as a first-class part of the methodology (Fig. 1) because the scaling decision in Eq. (1) and the choice to retain natural outliers rather than clip them were both informed directly by these diagnostics.

4.3 Logistic Regression

Logistic Regression models the probability of class membership using the sigmoid (logistic) function applied to a linear combination of input features:

$$P(y=1/x) = 1 / (1 + e^{-(wTx+b)}) \quad (2)$$

The model parameters w and b are estimated by minimizing the cross-entropy (log) loss over the training set of size N :

$$L(w,b) = -(1/N) \sum_i [y_i \log(p_i) + (1-y_i) \log(1-p_i)] \quad (3)$$

For the multi-class case (Normal / Suspect / Pathological), a one-vs-rest scheme extends the binary formulation across the three classes.

4.4 Support Vector Machine (SVM)

SVM seeks the hyperplane that maximizes the margin between classes. For a linearly separable case, the decision function is:

$$f(x) = \text{sign}(w^T x + b) \quad (4)$$

obtained by solving the constrained optimization problem:

$$\min_{\{w,b\}} \frac{1}{2} \|w\|^2 \text{ s.t. } y_i(w^T x_i + b) \geq 1 \quad \forall i \quad (5)$$

To handle the non-linear class boundaries present in CTG feature space, this study used the Radial Basis Function (RBF) kernel, which implicitly maps inputs into a higher-dimensional space:

$$K(x_i, x_j) = \exp(-\gamma \|x_i - x_j\|^2) \quad (6)$$

4.5 Random Forest

Random Forest is a bagging ensemble of decision trees, each trained on a bootstrap sample of the training data with a random subset of features considered at each split. Individual trees are grown by recursively selecting the split that minimizes Gini impurity:

$$G = 1 - \sum_k p_k^2 \quad (7)$$

where p_k is the proportion of class k samples in a node and C is the number of classes. Final predictions are obtained by majority vote across all T trees:

$$F(x) = \text{mode}\{h_1(x), h_2(x), \dots, h_T(x)\} \quad (8)$$

4.6 K-Nearest Neighbors (KNN)

KNN classifies a query point by majority vote among its k nearest neighbors in feature space, measured using Euclidean distance:

$$d(x, x') = \sqrt{\sum_i (x_i - x'_i)^2} \quad (9)$$

4.7 XGBoost

XGBoost is a regularized gradient boosting framework that builds an additive ensemble of decision trees sequentially, where each new tree f_t corrects the residual errors of the previous ensemble. The training objective at iteration t combines a differentiable loss function l with a regularization term Ω that penalizes tree complexity:

$$L^{(t)} = \sum_i l(y_i, F^{(t-1)}(x_i) + f_t(x_i)) + \Omega(f_t) \quad (10)$$

with the complexity penalty defined over the tree's leaf weights w and leaf count T_L :

$$\Omega(f) = \gamma T_L + \frac{1}{2} \lambda \sum_j w_j^2 \quad (11)$$

This regularization discourages overly complex trees, helping control overfitting — particularly important given the moderate size and class imbalance of the CTG dataset.

4.8 Ensemble Techniques: Blending and Stacking

Two ensemble strategies were constructed from the two strongest-performing base learners (XGBoost and Random Forest). The Blending (Soft Voting) Classifier averages the predicted class probabilities of its constituent models:

$$P(y=k/x) = (1/M) \sum_m P_m(y=k/x) \quad (12)$$

where M is the number of constituent models. The Stacking Classifier instead trains a meta-learner g on the out-of-fold predictions of the base learners:

$$F(x) = g(h_1(x), h_2(x), \dots, h_m(x)) \quad (13)$$

In this study, g is a Logistic Regression meta-learner trained on the predictions of XGBoost and Random Forest, using 5-fold cross-validation to generate out-of-fold base predictions and thereby reduce the risk of overfitting the meta-learner to the training set.

4.9 Evaluation Metrics

Given a confusion matrix with true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) per class, the following metrics were computed:

$$Accuracy = (TP + TN) / (TP + TN + FP + FN) \quad (14)$$

$$Precision = TP / (TP + FP) \quad Recall = TP / (TP + FN) \quad (15)$$

$$F1 = 2 \times (Precision \times Recall) / (Precision + Recall) \quad (16)$$

Because accuracy can be misleading under class imbalance, the Matthews Correlation Coefficient (MCC) was used as the primary comparison metric, as it accounts for all four confusion-matrix quantities symmetrically and yields a value in $[-1, 1]$, with 1 indicating perfect prediction:

$$MCC = (TP \times TN - FP \times FN) / \sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)} \quad (17)$$

ROC-AUC, the area under the receiver-operating-characteristic curve, was additionally computed in a one-vs-rest fashion for each class following Fawcett [10], providing a threshold-independent measure of class separability that complements the confusion-matrix-derived statistics above.

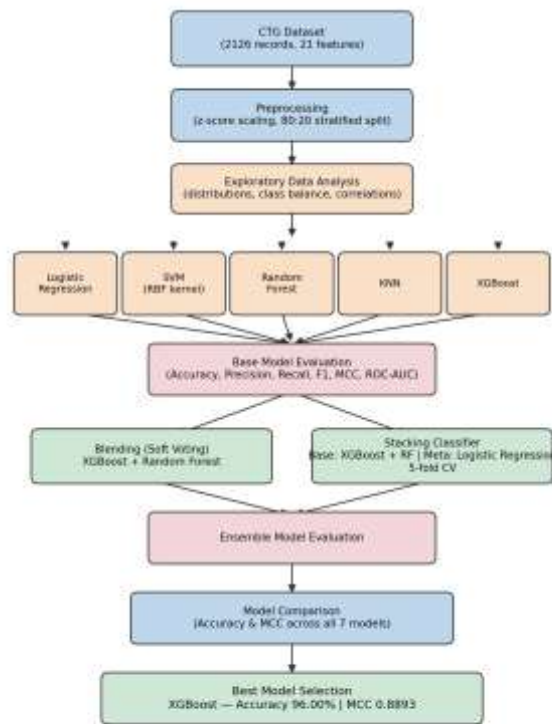


Fig. 1. End-to-end methodology for CTG-based fetal health classification.

5. Results and Discussion

Table 2 summarizes the accuracy and MCC achieved by each model on the held-out test set (20% of the data, $n = 426$).

Fig. 4 visualizes these results, and Fig. 5 shows per-class F1-score on the clinically critical Pathological class, computed directly from each model's confusion matrix (Fig. 6).

Model	Accuracy	MCC
Logistic Regression	0.8800	0.6655
SVM	0.9038	0.7302
Random Forest	0.9500	0.8474
KNN	0.9200	0.7575
XGBoost	0.9600	0.8893
Blending (Voting)	0.9577	0.8839
Stacking	0.9577	0.8829

Table 2. Model comparison — Accuracy and Matthews Correlation Coefficient (MCC).

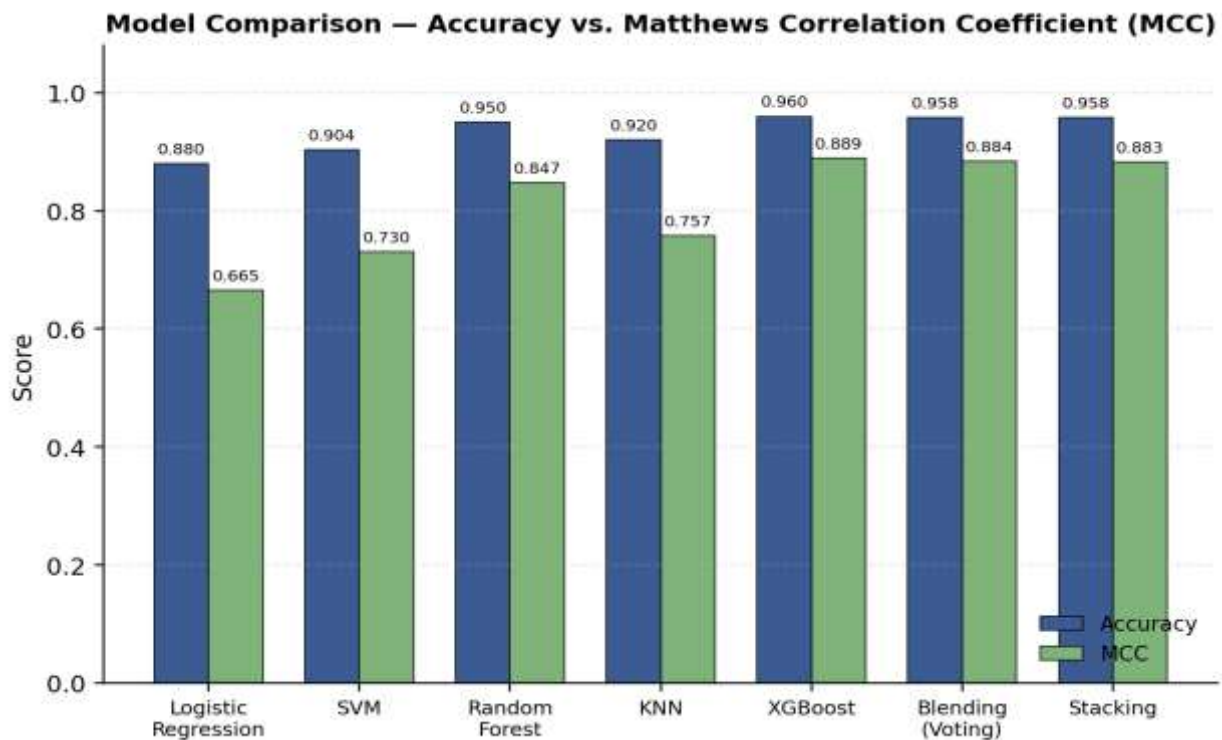


Fig. 4. Accuracy and MCC across all seven evaluated models.

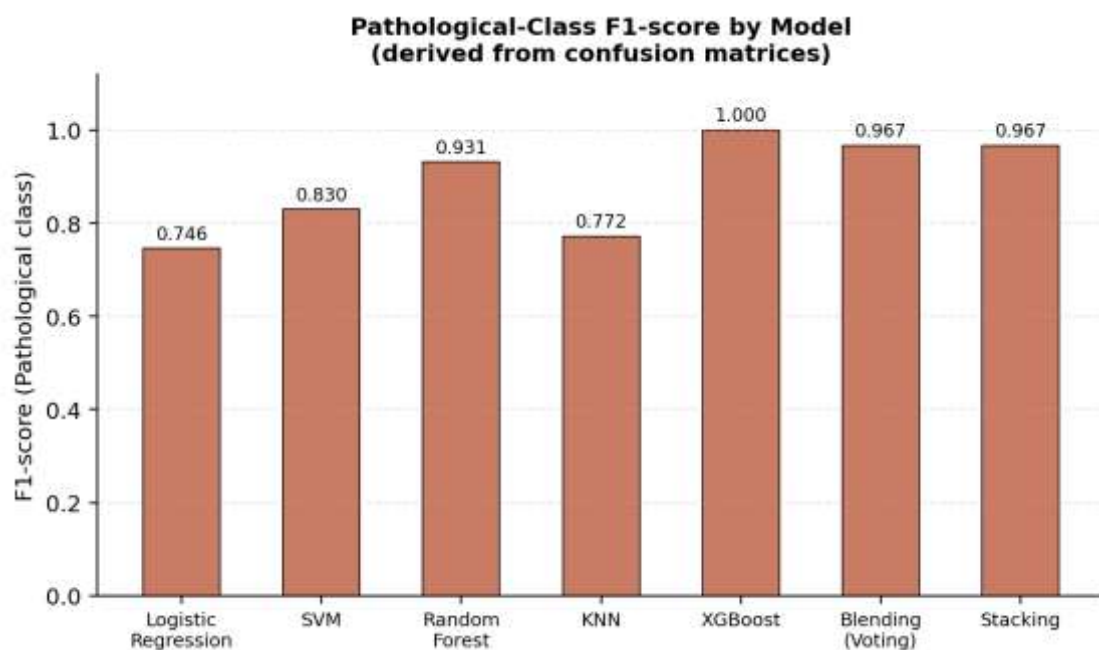


Fig. 5. F1-score on the Pathological class by model, computed from the confusion matrices in Fig. 6.

XGBoost achieved the highest standalone test accuracy (96.00%) and MCC (0.8893), narrowly outperforming both ensemble configurations (95.77% accuracy for each). Random Forest was the second-strongest base learner (95.00% accuracy, MCC 0.8474), while Logistic Regression, as expected for a linear baseline, achieved the lowest scores (88.00% accuracy, MCC 0.6655), reflecting its limited capacity to capture non-linear decision boundaries in the CTG feature space. Notably,

XGBoost achieved a perfect F1-score (1.000) on the minority Pathological class (Fig. 5), indicating complete sensitivity to the most clinically critical cases in this test split, followed closely by the Blending and Stacking ensembles (F1 = 0.967 each) and Random Forest (F1 = 0.931).

5.1 Confusion Matrix Analysis

Confusion Matrices for All Seven Evaluated Models (Test set, n = 426)

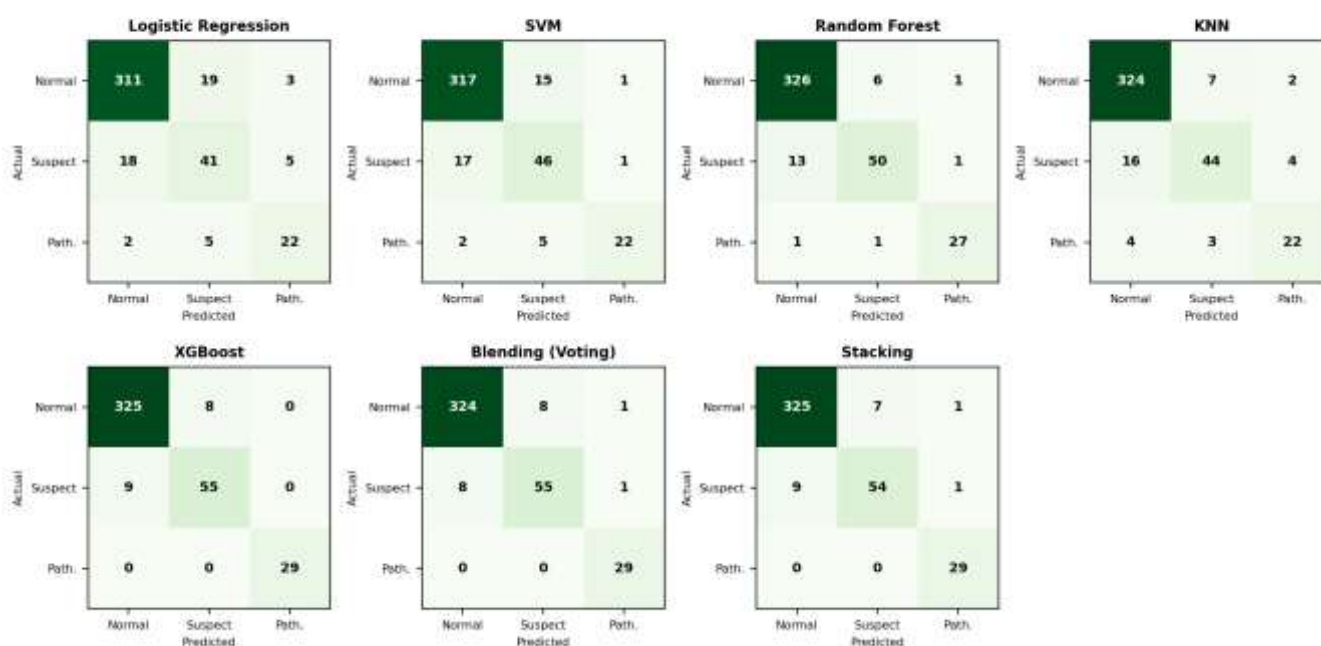


Fig. 6. Confusion matrices for all seven models (rows/columns: Normal, Suspect, Pathological; test set, n = 426).

The confusion matrices in Fig. 6 reveal that most misclassification for every model occurs between the Normal and Suspect classes, which are clinically the most similar and hardest to separate from summary CTG statistics alone. Misclassification into or out of the Pathological class is rare across all models, and is essentially eliminated for XGBoost, Blending, and Stacking, which each mislabel zero Pathological cases as Normal — a particularly reassuring property from a patient-safety perspective, since a false-negative Pathological prediction (i.e., a truly distressed fetus classified as Normal) is the most clinically costly type of error in this setting.

Table 3 breaks down precision, recall, and F1-score per class for the best-performing model (XGBoost), computed directly from its confusion matrix in Fig. 6. The Suspect class is the hardest for every model, including XGBoost, to identify precisely (precision 0.873) — consistent with it being a clinical borderline category that shares boundary characteristics with both Normal and Pathological cases — while Normal and Pathological are both classified with precision and recall at or above 0.97.

Class	Prec.	Recall	F1
Normal	0.973	0.976	0.974
Suspect	0.873	0.859	0.866
Path.	1.000	1.000	1.000

Table 3. Per-class precision, recall, and F1-score for the best model (XGBoost), derived from its confusion matrix in Fig. 6.

5.2 Cross-Validation Stability

The Stacking classifier's 5-fold cross-validation on the training set yielded a mean accuracy of 94.29% (fold range 92.94%–95.59%), suggesting stable generalization despite not exceeding XGBoost's single-model test performance. This narrow spread indicates the ensemble's performance is not an artifact of a favorable train-test split, supporting its practical reliability even where it does not top the leaderboard on this particular test set.

Overall, these results suggest that, on this dataset and feature set, a well-tuned gradient boosting model alone can match or exceed the benefit typically expected from ensembling with weaker base learners, though ensemble methods may still offer improved robustness across different data samples and are valuable where minimizing variance in deployment is a priority.

5.3 Bias–Variance Perspective and Computational Considerations

The ordering of model performance observed here is broadly consistent with the bias–variance characteristics of each algorithm family. Logistic Regression, as a single global linear decision boundary, exhibits high bias on a dataset whose

class boundaries are visibly non-linear in the EDA pairplots, which explains both its lowest accuracy and, more tellingly, its steep drop in MCC (0.6655) relative to accuracy (0.8800) — a signature of a classifier that performs well on the majority class but struggles disproportionately on the minority classes. KNN, despite being a non-linear, non-parametric method, is sensitive to the local density imbalance between classes: because Normal points vastly outnumber Pathological points in feature space, a query near the Suspect/Pathological boundary is more likely to be surrounded by Normal neighbors, biasing predictions toward the majority class even after scaling. Random Forest and XGBoost, by contrast, partition the feature space recursively and can isolate small, locally homogeneous regions corresponding to the minority classes, which is reflected in their near-perfect Pathological-class F1-scores (0.931 and 1.000, respectively).

From a deployment perspective, inference cost also differs meaningfully across the seven models. Logistic Regression and the trained SVM require only a single forward pass through a linear (or kernel) decision function and are the cheapest to score at inference time. KNN requires a distance computation against the full training set at prediction time, making it the least scalable of the base learners as data volume grows, though this is a minor concern at the present dataset size ($n = 2126$). Random Forest and XGBoost both require traversing multiple trees per prediction, but modern gradient boosting implementations are heavily optimized and, in practice, add only marginal latency over a single tree. The two ensemble methods necessarily incur the combined cost of their constituent base learners (XGBoost and Random Forest) plus, for Stacking, a lightweight Logistic Regression meta-learner pass; given the ensembles did not outperform standalone XGBoost here, this additional cost is not currently justified purely on accuracy or MCC grounds, though it may still be worthwhile where minimizing variance in deployment — rather than raw peak accuracy — is the priority, as discussed in Section 5.2.

6. Limitations and Future Work

This study has several limitations. First, all models were evaluated on a single train-test split from one publicly available dataset; external validation on independent CTG datasets from different hospitals and populations is needed to confirm generalizability, since the 2126 recordings used here were collected at a single center and annotated by a fixed panel of three obstetricians, whose labeling conventions may not transfer perfectly to other institutions. Second, default or lightly tuned hyperparameters were used for each classifier; systematic hyperparameter optimization (e.g., grid search, Bayesian optimization, or successive-halving search) could further improve performance, particularly for SVM and KNN, which are sensitive to kernel and neighborhood-size choices respectively. Third, while MCC was used to address class imbalance at the evaluation level, explicit resampling techniques such as SMOTE [15] were not applied at the training level and could be explored in future work, potentially in combination with cost-sensitive loss weighting so that Pathological-class errors are penalized more heavily during

training rather than only measured more carefully at evaluation time.

Fourth, the ground-truth labels themselves are the product of expert consensus rather than a fully objective gold standard, and the inter-observer variability discussed in Section 1 that motivates automated classification in the first place also, unavoidably, propagates some label noise into the training data; no amount of model benchmarking can fully correct for this upstream uncertainty. Fifth, this study reports point estimates of test-set performance rather than confidence intervals or statistical significance tests (e.g., McNemar's test or a paired bootstrap) between the top-performing models; given that XGBoost, Blending, and Stacking differ by only 1–2 misclassified cases out of 426, a formal significance test would help determine whether XGBoost's narrow lead over the ensembles reflects a genuine algorithmic advantage or sampling noise in this particular split. Finally, model interpretability techniques such as SHAP or LIME would help clinicians understand individual predictions and build trust in a deployed system, and are a natural next step toward clinical translation.

Future work should therefore prioritize, in rough order of expected impact: (i) external validation on multi-center CTG datasets; (ii) paired statistical significance testing between the top three models; (iii) systematic hyperparameter search combined with nested cross-validation to avoid optimistic bias; (iv) exploration of class-imbalance-aware training strategies such as SMOTE or focal loss; and (v) integration of post-hoc interpretability tooling to support clinical adoption.

7. Conclusion

This study benchmarked five machine learning classifiers and two ensemble techniques for fetal health classification using cardiotocogram data, grounding each method in its underlying mathematical formulation and evaluating performance using accuracy, precision, recall, F1-score, confusion matrices, and the Matthews Correlation Coefficient. XGBoost emerged as the most accurate and reliable standalone model, achieving 96.00% accuracy and an MCC of 0.8893, slightly ahead of the Blending and Stacking ensembles, and achieved perfect sensitivity on the clinically critical Pathological class. These findings highlight the strength of regularized gradient boosting for structured, imbalanced clinical datasets and support its potential integration into CTG-based clinical decision-support systems. Future work should explore systematic hyperparameter optimization, external validation on independent CTG datasets, class-imbalance-aware resampling, and interpretability methods to increase clinical trust and support real-world adoption.

Acknowledgment

The authors thank the Department of Computer Science and Engineering, BITS Bhopal, for institutional support, and the organizers of ICCSAM 2026 for the opportunity to present this work.

References

- [1] World Health Organization, "Stillbirth," WHO Fact Sheets, Geneva, Switzerland, 2023. [Online]. Available: <https://www.who.int/news-room/fact-sheets/detail/stillbirth>
- [2] C. Cortes and V. Vapnik, "Support-vector networks," Machine Learning, vol. 20, no. 3, pp. 273–297, 1995.
- [3] T. Cover and P. Hart, "Nearest neighbor pattern classification," IEEE Transactions on Information Theory, vol. 13, no. 1, pp. 21–27, 1967.
- [4] L. Breiman, "Random forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.
- [5] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, Classification and Regression Trees. Belmont, CA: Wadsworth International Group, 1984.
- [6] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in Proc. 22nd ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, 2016, pp. 785–794.
- [7] D. H. Wolpert, "Stacked generalization," Neural Networks, vol. 5, no. 2, pp. 241–259, 1992.
- [8] T. G. Dietterich, "Ensemble methods in machine learning," in Multiple Classifier Systems, LNCS vol. 1857, 2000, pp. 1–15.
- [9] B. W. Matthews, "Comparison of the predicted and observed secondary structure of T4 phage lysozyme," Biochimica et Biophysica Acta, vol. 405, no. 2, pp. 442–451, 1975.
- [10] T. Fawcett, "An introduction to ROC analysis," Pattern Recognition Letters, vol. 27, no. 8, pp. 861–874, 2006.
- [11] D. R. Cox, "The regression analysis of binary sequences," Journal of the Royal Statistical Society: Series B, vol. 20, no. 2, pp. 215–242, 1958.
- [12] D. Ayres-de-Campos, J. Bernardes, A. Garrido, J. Marques-de-Sá, and L. Pereira-Leite, "SisPorto 2.0: A program for automated analysis of cardiotocograms," Journal of Maternal-Fetal Medicine, vol. 9, no. 5, pp. 311–318, 2000.
- [13] D. Campos and J. Bernardes, "Cardiotocography [Dataset]," UCI Machine Learning Repository, 2000. [Online]. Available: <https://doi.org/10.24432/C51S4N>
- [14] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.
- [15] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," Journal of Artificial Intelligence Research, vol. 16, pp. 321–357, 2002.
- [16] D. Chicco and G. Jurman, "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation," BMC Genomics, vol. 21, no. 1, art. 6, 2020.
- [17] M. Piri, M. R. Delavar, and F. Sadeghi-Niaraki, "An improved model ensembled of different hyper-parameter tuned machine learning algorithms for fetal health prediction," arXiv preprint arXiv:2305.17156, 2023.