Fetal Health Classification

An Exploration of Various Methods

Agathe Benichou
M.Sc. in Machine Learning & Data Science
Reichman University, Department of Computer Science

In collaboration with Nadav Loebl, Head of AI at Beilinson Innovation (Beilinson Hospital)

This paper addresses the challenge of accurately classifying fetal health based on cardiotocogram exam signals. The methodology involves exploring various machine learning models, including Logistic Regression, Decision Trees, XGBoost and Voting Classifiers. The results demonstrate that ensemble models achieve higher sensitivity and minimize false negatives. The implications of this research highlight the potential for improved fetal health monitoring and reducing unnecessary interventions.

Introduction

Background

Fetal health monitoring is a critical aspect of prenatal care and is essential for ensuring successful child birth. Reduction of child mortality is a key indicator of human progress, as is reflected in the United Nations' Sustainable Development Goals. The UN aims to end preventable deaths of newborns and children under five years of age by 2030, with all countries working to reduce under-five mortality to at least 25 per 1,000 live births. Maternal mortality, much like child mortality, results in approximately 295,000 deaths during or after pregnancy and childbirth, with the majority occurring in low-resource settings. Most of these deaths could be prevented with timely and accurate medical interventions (Maulik & Mundy, 2000).

Cardiotocograms (CTGs) provide a simple and cost-effective option to assess fetal health and allows healthcare professionals to take preventive action to avoid child and/or maternal mortality. CTGs work by sending ultrasound pulses and reading the responses. It provides information on fetal heart rate (FHR), fetal movements, uterine contractions, and more. CTG is a form of Doppler that uses only sound and does not produce an image. It is a simple, painless and non-invasive procedure performed before birth and during labor to monitor the baby for any signs of distress. Two electronic sensors are strapped to the mother's belly: one monitors the heartbeat and movement, and the other records uterine contractions (see Figure 1). These sensors are connected to a machine with elastic belts holding them in place. CTG monitors the baby's heartbeat while moving and resting to see if the heart rate increases with movement, which indirectly indicates sufficient oxygen from the placenta. Doctors check if the test result is reactive (the baby's heart rate increases by the expected amount after each movement) or non-reactive (the heart rate does not increase) which could indicate an issue (Bailey, 2009).

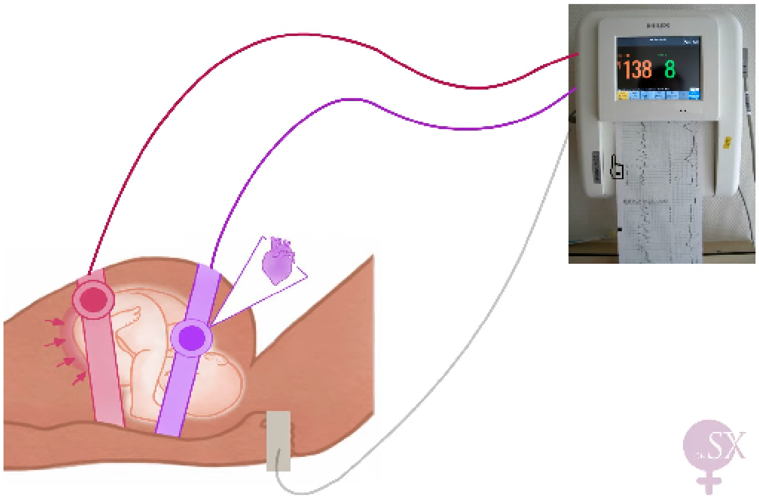


Figure 1: View of how CTG exams work

In Figure 1, the purple line represents the sensor that monitors the heart rate of the fetus. The pink line is the sensor that measures uterine contractions by assessing the tension of the abdominal wall. (Bailey, 2009).

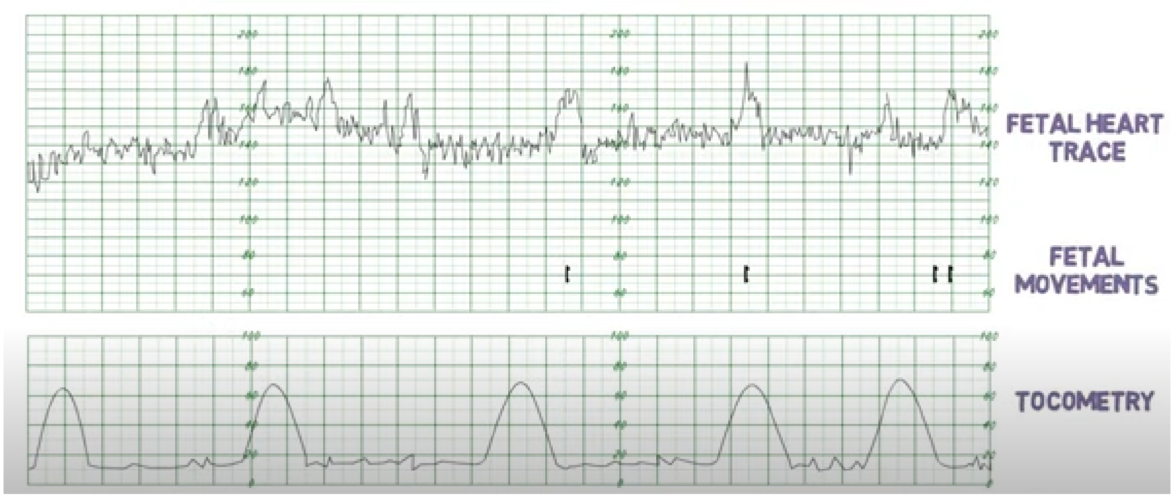


Figure 2: Fetal heart trace, fetal, and uterine contractions over a ten-minute period

In Figure 2, the upper section (fetal heart trace) shows the fetal heart rate, the middle section (fetal movements) represents fetal movements and the bottom section, label (Tocometry) records uterine contractions. Each box on the graph represents a duration of one minute, providing a detailed view of the fetal and maternal status during monitoring. The total time span covered in this chart is ten minutes, offering a comprehensive overview during this period (Bailey, 2009).

CTG interpretation is vital for identifying fetal distress and ensuring timely intervention. Medical professionals use the DR C BRAVADO method for CTG assessment. This method includes:

- **DR (Defined Risk)**: Assesses previous risk factors of pregnancy.
- **C (Contractions)**: Evaluates the frequency and duration of uterine contractions.
- **BRA (Baseline Rate)**: Monitors the average heart rate of the fetus..
- **V (Variability)**: Measures beat-to-beat variability in FHR.
- **A (Accelerations)**: Checks for increases in FHR by more than 15 BPM for more than 15 seconds, indicating healthy fetal movements.
- **D (Decelerations)**: Monitors decrease in FHR by more than 15 BPM for more than 15 seconds. Decelerations can be early, variable, or late, each with different implications.

CTG patterns are categorized into normal, suspect, and pathological based on baseline rate, variability, and decelerations. This categorization helps in deciding the necessary interventions (Bailey, 2009).

Problem Statement

This research focuses on developing a machine learning model to classify fetal health status into normal, suspect, and pathological categories using CTG signals, with the goal of accurately identifying fetal distress and improving labor outcomes. Accurate classification is essential to prevent adverse neonatal outcomes and reduce both child and maternal mortality by minimizing misclassification, which can lead to unnecessary interventions or missed diagnoses. The study will evaluate the performance of different models to determine the most effective approach.

Related Works

Recent studies have examined the potential of machine learning to predict fetal health status using CTG data. In one study, several machine learning models were applied, including k-Nearest Neighbors, Support Vector Machines, Random Forest, Gradient Boosting, and Neural Networks, to classify fetal health into different categories. Among these, the Random Forest model demonstrated the highest accuracy of 95.77%. The research underscored the importance of factors such as the frequency of accelerations and variability in fetal heart rate as key predictors of fetal health outcomes (Dixit, 2021).

Another study highlighted the critical role of feature engineering and addressing class imbalances in enhancing model performance. To balance the dataset and improve predictive accuracy, techniques like SMOTE were employed. (Liu & Zhang, 2023).

Several limitations have been identified when classifying fetal health status. One key issue is the imbalanced datasets, where normal cases far outnumber suspect and pathological cases and this leads to biased predictions. Many studies rely on traditional feature selection methods, which can limit model accuracy. Generalization is also a concern, as studies often use small, localized datasets that reduce the models' applicability to other populations. The complexity of certain machine learning models, such as neural networks, makes them difficult to interpret, hindering their integration into clinical practice.

Data

The dataset was obtained from Kaggle (Andrew Mvd, 2020), consisting of 2,126 records from CTG exams. It consists of 22 columns, each representing various features extracted from CTG exams. Each row in the dataset represents a different patient. The features include:

- **Baseline Value:** Baseline FHR, ranging from 106 to 160 BPM. The normal range is between 110 and 160 BPM; values below 110 BPM could indicate fetal distress, and values above 160 BPM could indicate fetal infections.
- **Accelerations:** Number of accelerations per second, ranging from 0 to 0.02. Accelerations are short-term rises in the fetal heart rate.
- **Fetal Movement:** Number of fetal movements per second, ranging from 0 to 0.48.
- **Uterine Contractions:** Number of uterine contractions per second, ranging from 0 to 0.01.
- **Light Decelerations:** Number of light decelerations per second, ranging from 0 to 0.01.
- **Severe Decelerations:** Number of severe decelerations per second, ranging from 0 to 0.01.
- **Prolonged Decelerations:** Number of prolonged decelerations per second, ranging from 0 to 0.01.
- **Abnormal Short Term Variability:** Percentage of time with abnormal short-term variability, ranging from 12 to 87. This refers to the reduction or absence of fluctuations in the fetal heart rate over a defined period.
- **Mean Value of Short Term Variability:** Mean value of short-term variability, ranging from 0.2 to 7.
- **Percentage of Time with Abnormal Long Term Variability:** Percentage of time with abnormal long-term variability, ranging from 0 to 91.
- **Mean Value of Long Term Variability:** Mean value of long-term variability, ranging from 0 to 50.7.
- **Histogram Width**: This represents the range of fetal heart rate (FHR) values observed during the CTG exam. A wider histogram width (ranging from 3 to 180) indicates a broader range of heart rate values, suggesting more variability in the fetal heart rate.
- **Histogram Min**: The minimum value of the histogram (ranging from 50 to 159) represents the lowest FHR value observed during the CTG exam, indicating the minimum heart rate recorded.
- **Histogram Max**: The maximum value of the histogram (ranging from 122 to 238) represents the highest FHR value observed during the CTG exam, indicating the maximum heart rate recorded.
- **Histogram Number of Peaks**: This counts the number of peaks (ranging from 0 to 18) in the histogram, which correspond to the most frequently occurring FHR values. More peaks can indicate multiple common heart rate levels during the exam.
- **Histogram Number of Zeroes**: This counts the number of times the FHR was recorded as zero (ranging from 0 to 10), which may indicate periods of no detectable heartbeat or signal loss.
- **Histogram Mode**: The mode value of the histogram (ranging from 60 to 187) represents the most frequently occurring FHR value during the exam. It shows the heart rate value that appeared most often.

- **Histogram Mean**: The mean value of the histogram (ranging from 73 to 182) is the average of all the FHR values during the CTG exam, providing a central tendency of the heart rate data.
- **Histogram Median**: The median value of the histogram (ranging from 77 to 186) represents the middle value of the FHR data when it is sorted in order. Half of the heart rate values are below this number, and half are above it.
- **Histogram Variance**: The variance of the histogram (ranging from 0 to 269) measures the spread or variability of the FHR values. A higher variance indicates that the heart rate values are more spread out, showing greater fluctuation.
- **Histogram Trend**: The trend of the histogram (with values -1, 0, or 1) indicates the overall direction of the FHR data over time. A trend value of -1 indicates a downward trend, 0 indicates no trend, and 1 indicates an upward trend.
- **Fetal Health: Classification** of fetal health status into normal (labeled 1), suspect (labeled 2), and pathological (labeled 3) categories.

While the dataset captures essential features of a CTG, it lacks several important elements that could enhance diagnostic accuracy. The dataset classifies different types of decelerations)light, severe, prolonged) but does not specify the duration of decelerations is missing. Baseline variability is needed to distinguish between short-term and long-term variability and is also absent. Most importantly, contextual factors such as maternal health, medication, gestational age, and pre-existing conditions are not provided. The dataset omits recovery time (the time taken for the fetal heart rate to return to baseline) and external factors like whether the mother consumed sugary substances beforehand and the total duration of monitoring. Other missing elements include previous CTG results and additional obstetric measures such as fetal blood vessels and amniotic fluid index, all of which could provide a more comprehensive understanding of fetal health. Furthermore, the dataset does not specify whether the CTG was conducted during labor or a routine check-up, nor does it include information on the maternal position during the CTG acquisition.

Data Assumptions

We assume that each row in the dataset represents an individual CTG recording from different mothers during a particular observation period. Each column is assumed to reflect features extracted from the CTG exams, averaged across the entire exam.

A key limitation of this dataset is the lack of a defined time frame for each CTG exam. FHR patterns and their clinical interpretations are highly time-sensitive. Without temporal data, it is difficult to fully utilize the predictive capabilities of CTGs for monitoring fetal health. Therefore, we assume a 10-minute recording interval because it aligns with standard practice (Ayres-de-Campos et al., 2000).

Data Observations

The dataset is imbalanced, with a significant majority of the records classified as Normal (labeled 1) and the rest split to Suspect (labeled 2) and Pathological (labeled 3) (see Figure 3).

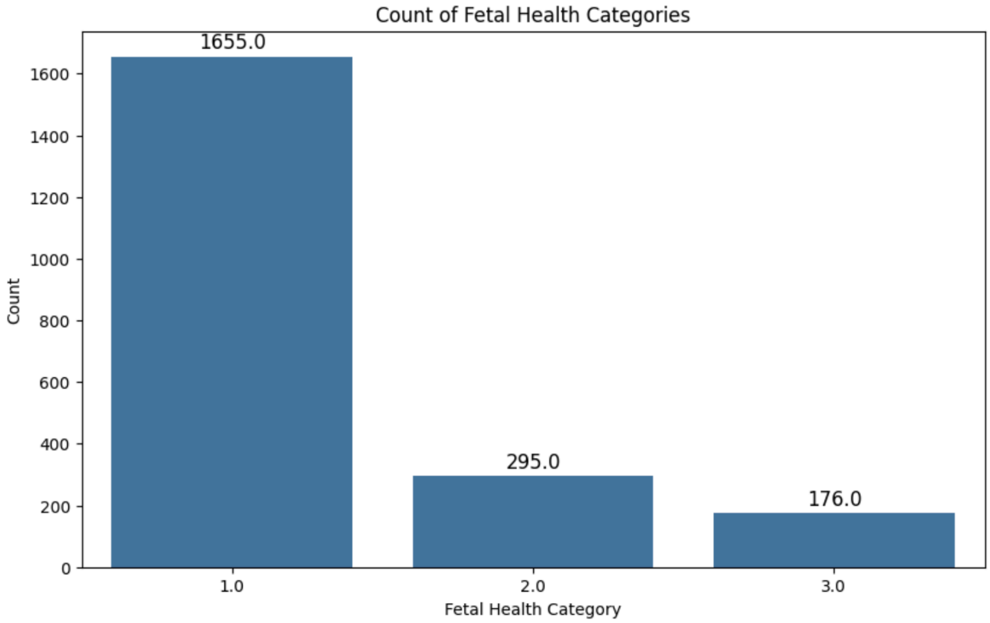


Figure 3: Bar chart showing the distribution of fetal health categories

The box plot in Figure 4 shows the distribution of baseline FHR across the three fetal health categories. For the normal category, the baseline FHR has a wider range with a median around 130 BPM. The suspect category shows a higher median baseline FHR around 140 BPM and a larger interquartile range. The pathological category has the lowest median baseline FHR around 120 BPM and a narrower range.

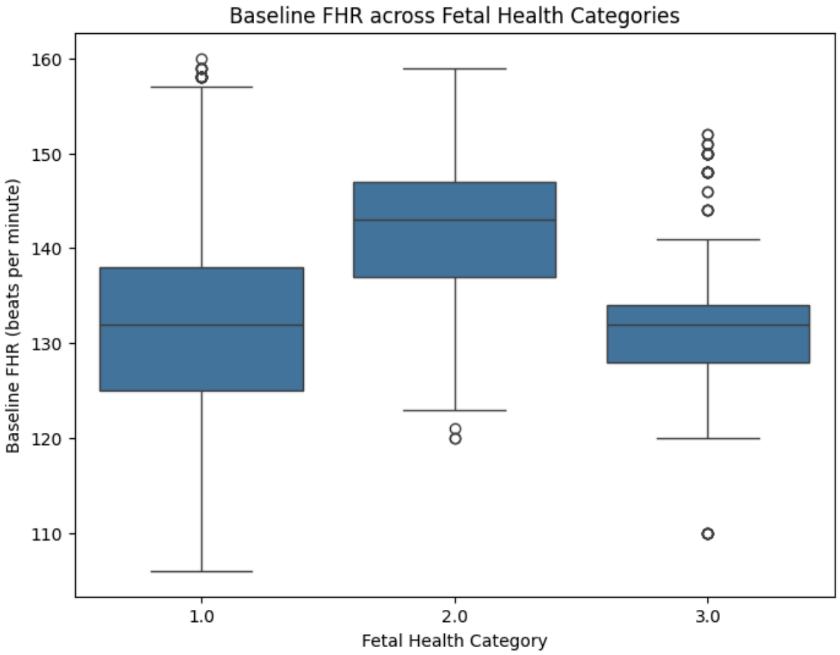


Figure 4: Box plot showing the distribution of baseline FHR across the fetal health categories

The correlation matrix (Figure 5) provides a view of the relationships between different features. There are strong positive correlations between the histogram-related features, indicating that these features tend to vary together. There are some moderate negative correlations, such as between baseline_value and histogram_min. The fetal_health has varying degrees of correlation with different features.

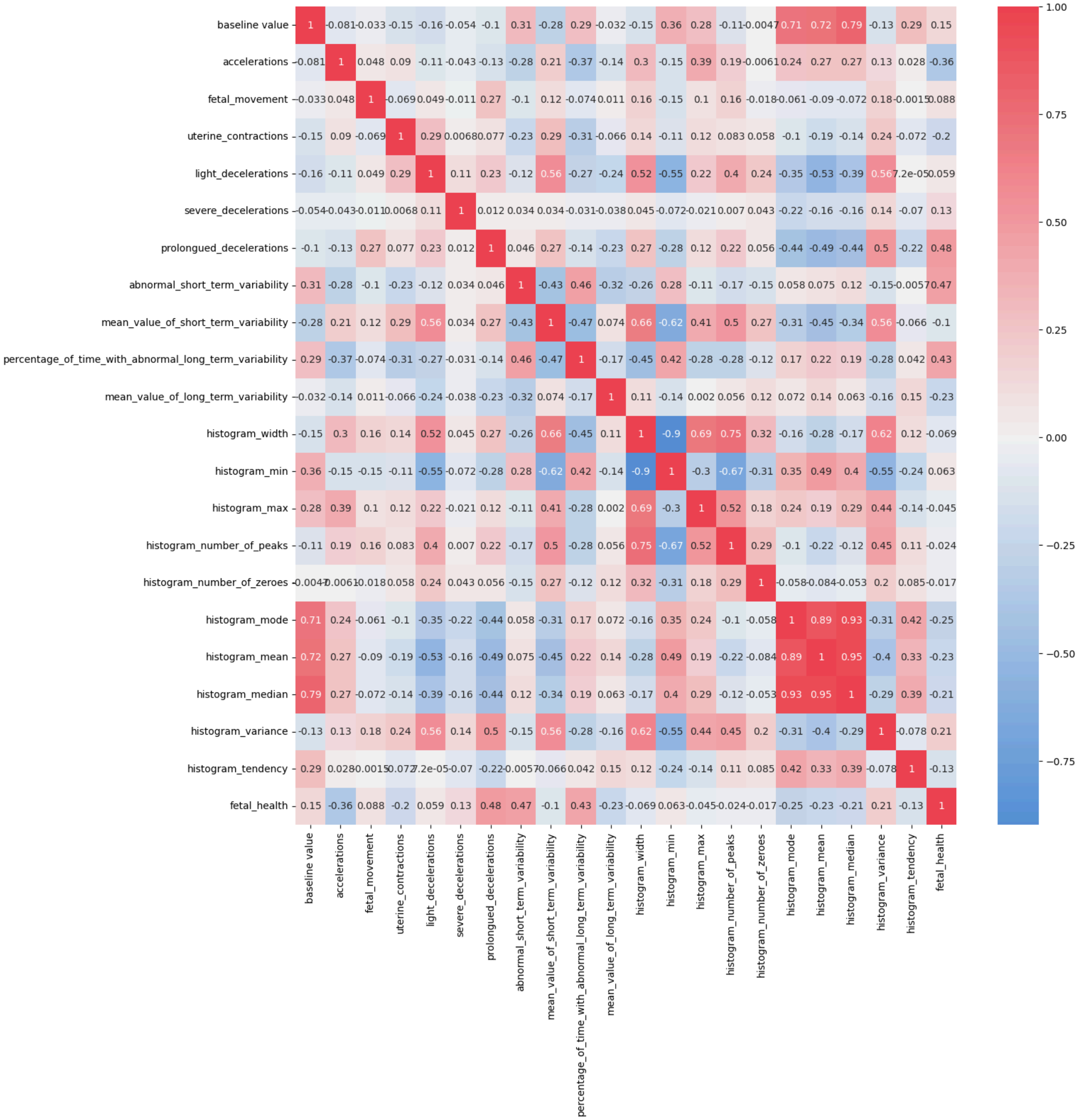


Figure 5: Correlation matrix showing the relationships between features in the dataset.

This correlation matrix (Figure 6) offers specific insights into the relationships between the features. There are positive correlations, such as between prolonged_decelerations and fetal_health, indicating that an increase in prolonged decelerations is associated with a higher likelihood of fetal health issues. Accelerations show a strong negative correlation with fetal health, suggesting that fewer accelerations are associated with worse fetal health outcomes.

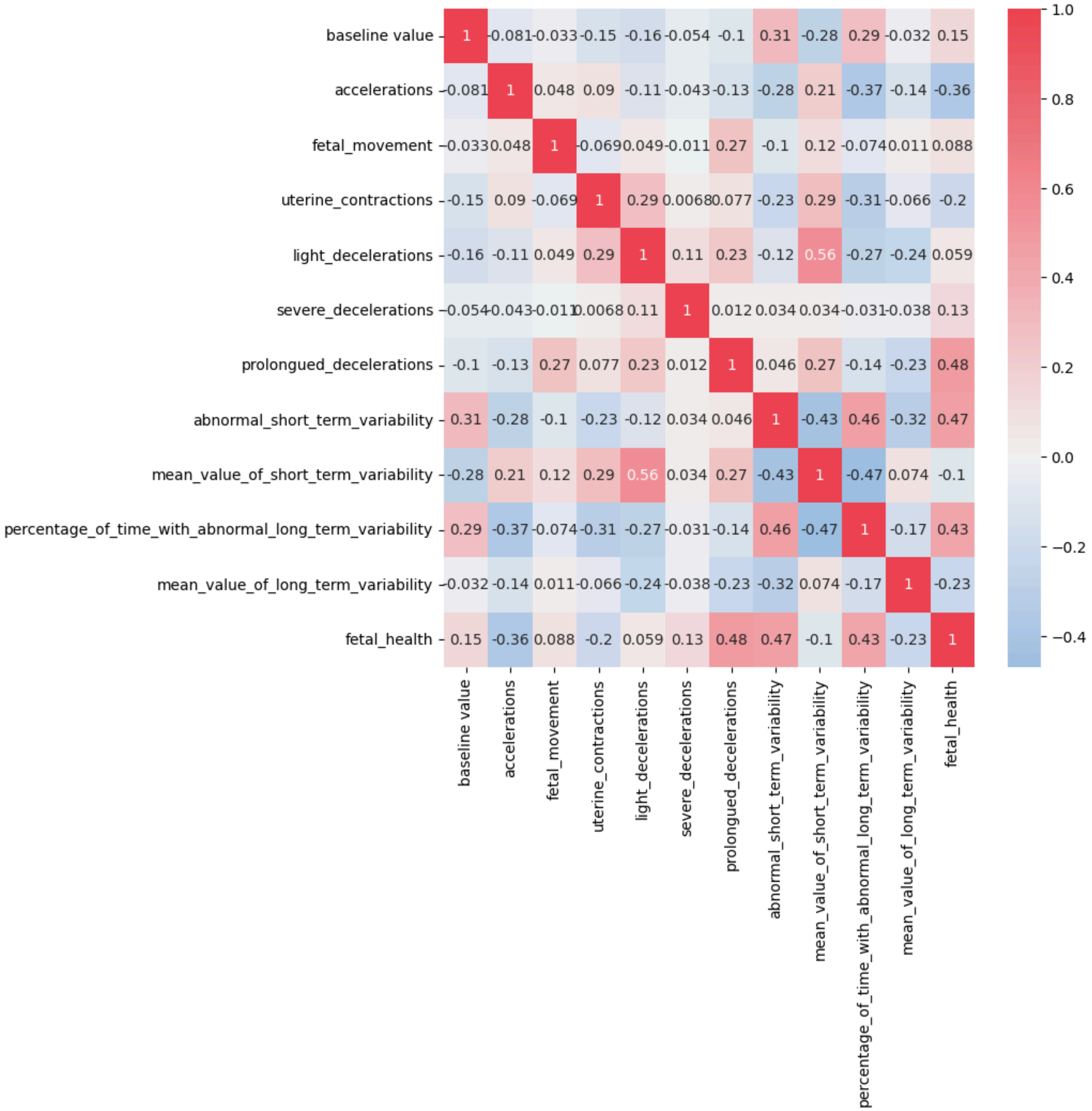


Figure 6: Updated correlation matrix highlighting the relationships between features and fetal health categories

The histogram of baseline FHR (Figure 7) shows a roughly normal distribution centered around 130 beats per minute. This can help in identifying abnormal values that might indicate fetal distress.

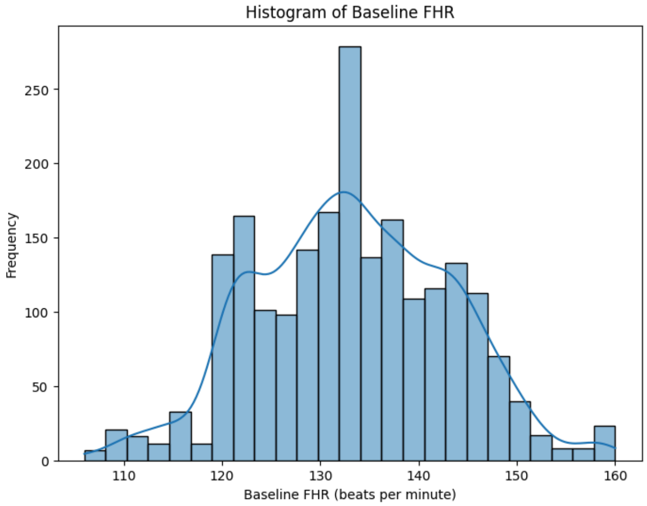


Figure 7: Histogram showing the distribution of baseline fetal heart rate (FHR) across all records

The pair plot of prolonged decelerations and abnormal short-term variability in Figure 8 shows higher concentrations in the suspect and pathological categories, indicating these features are strong indicators of adverse fetal health. The density plots support this, with suspect and pathological categories displaying higher values compared to the normal category.

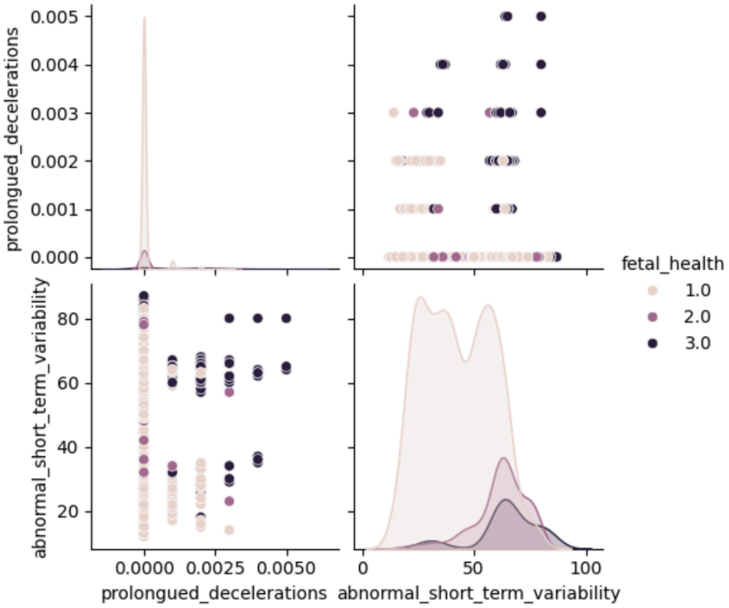


Figure 8: Pair plot of prolonged decelerations and short-term variability

To simulate FHR data, histogram features like minimum, maximum, mode, mean, median, and variance of the FHR were used to generate a realistic pattern. A distribution was centered around the mean FHR with the variance. The simulation generated random data points following a normal distribution based on the mean and variance values. This distribution was adjusted to ensure the FHR values remained within the specified minimum and maximum range, so no extreme outliers. Peaks and zeroes were introduced based on the frequency observed in the original dataset, and any overall trend in the FHR over time (increasing, decreasing, or stable trends) was considered by incorporating gradual shifts in the baseline. This simulated data offers a visual representation of what a typical 10-minute recording might look like (see Figure 9).

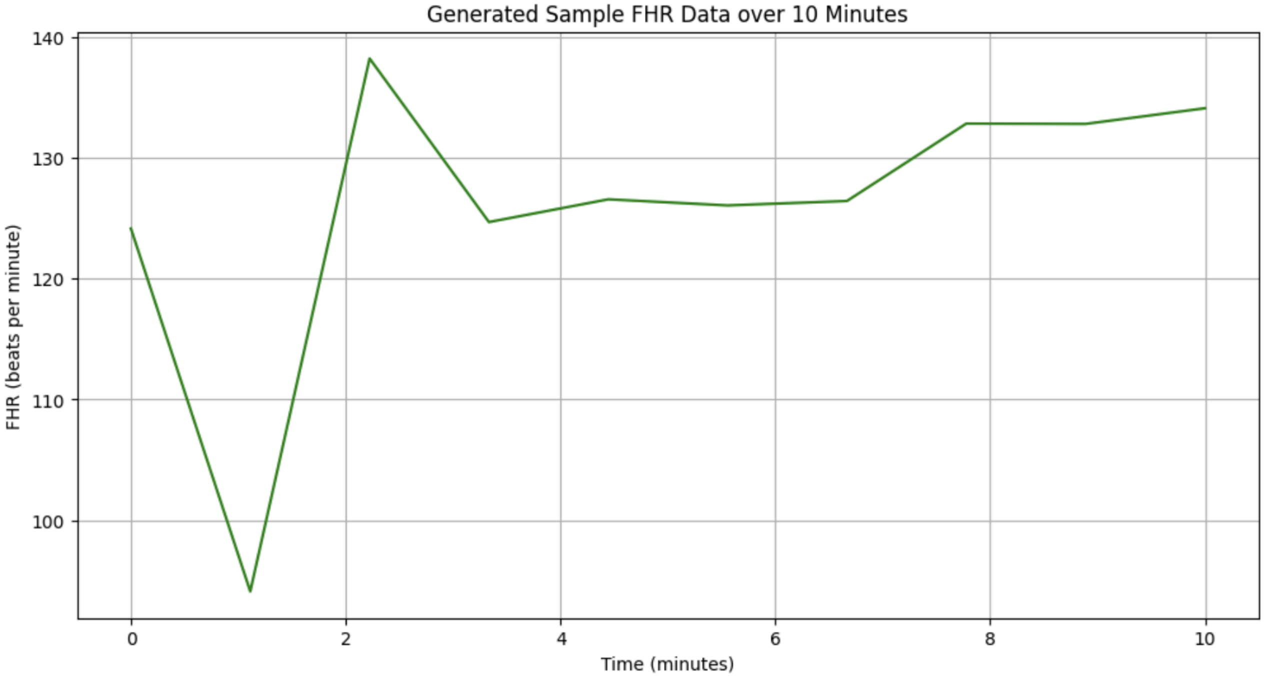


Figure 9: Simulated FHR data over 10 minutes.

The generated FHR samples for different health categories were created using distinct patterns for normal, suspect, and pathological cases. For each health category, specific variance and trend modifications were applied to reflect typical behavior. For instance, normal health patterns displayed a stable FHR with minor fluctuations around the mean, while suspect health patterns exhibited moderate fluctuations with occasional decelerations or accelerations. Pathological health patterns showed significant variability, including larger deviations from the mean and frequent decelerations. The generated data for each category was plotted over a 30-minute period to highlight these trends. This helps in understanding typical FHR patterns associated with each fetal health status (see Figure 10).

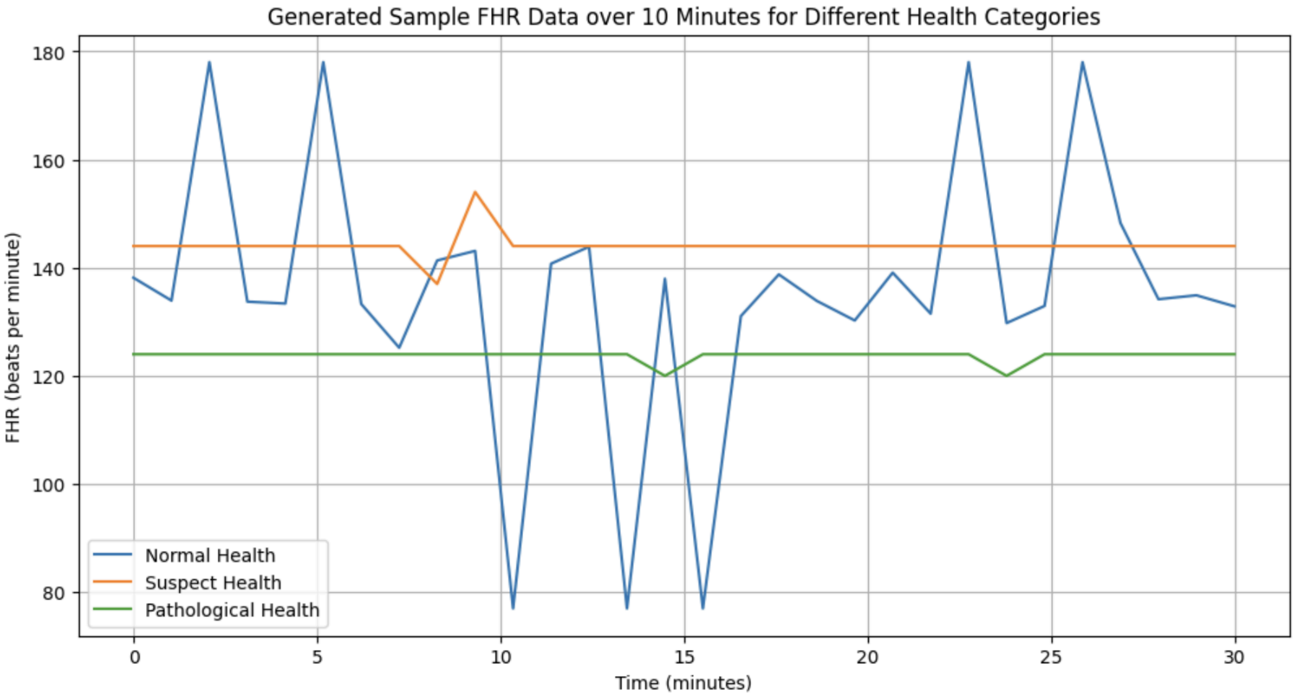


Figure 10: Simulated FHR data for different health categories over 30 minutes.

Features such as fetal heart rate, prolonged decelerations, and abnormal short-term variability show significant correlations with adverse fetal outcomes. Prolonged decelerations have a strong positive correlation with the pathological class. The histogram-related features offer insights into the variability of the FHR during the CTG exams. These features have a negative correlation with fetal health issues and are critical in assessing the overall health and movement patterns of the fetus.

Methodology

The methodology began with establishing a baseline using models such as Logistic Regression and Decision Trees, to provide foundational insights into the data. Feature selection techniques were employed to identify the most significant features. To address class imbalances, SMOTE was applied to generate synthetic samples for minority classes.

Following the baseline evaluation, more sophisticated models, including Support Vector Machines (SVM), XGBoost, and CatBoost, were explored. Hyperparameter tuning and optimization techniques were applied to fine-tune these models for optimal performance across the different fetal health classes.

Once the best-performing model for each class was identified, these models were integrated into ensemble models to capitalize on their individual strengths. Two different ensemble strategies were compared: one focused on maximizing overall accuracy and the other employed a voting mechanism. The final selection of the ensemble model was based on its ability to provide a balanced and robust performance across all fetal health categories.

Data Preprocessing

In the data preprocessing phase, the dataset was first divided into features and the target variable, fetal_health, which classifies fetal health status into Normal, Suspect, and Pathological categories. The data was initially imported once, and a universal test set was created at the outset to ensure that each baseline model was evaluated on a consistent and completely unseen set of data. The dataset was then split into training (70% or 1,488 samples), validation (15% or 319 samples), and test (15% or 319 samples) sets using stratified sampling to maintain the class distributions across all subsets.

To prepare the data for modeling, input features were standardized, giving each feature a mean of 0 and a standard deviation of 1. The target labels in the validation and test sets were binarized for specific model evaluations. Stratified sampling was employed throughout the splitting process to preserve the original class distribution in all subsets, and each model was evaluated on the same test set to provide a fair and consistent comparison of performance across all models.

Evaluation Metrics

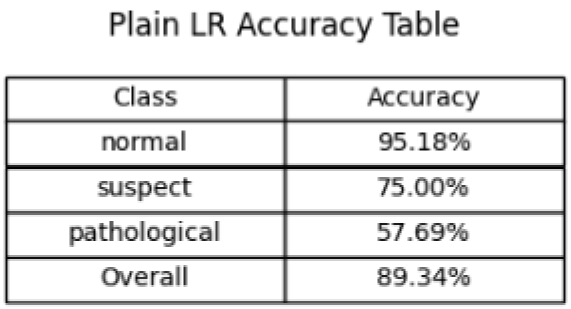
Model performance was comprehensively evaluated using metrics such as accuracy, precision, recall and AUC:

- **Accuracy:** Measures the overall correctness of the model by calculating the proportion of true results (both true positives and true negatives) among the total number of cases. It provides an overview of how well the model is performing across all classes.
- **Precision (Positive Predictive Value, PPV):** Measures out of all the true positives, how many were actually caught. Precision is crucial in reducing false positives, which is important for medical contexts where misdiagnosing healthy fetuses can lead to unnecessary interventions.
- **Recall (Sensitivity):** Measures how many actual positive cases were correctly identified. High sensitivity ensures that positive cases are not missed.
- **Area Under the Curve (AUC):** Assesses the model's ability to distinguish between classes. AUC represents the degree of separability, indicating how well the model can distinguish between classes. Useful in medical diagnosis because it balances sensitivity and specificity.
- **Negative Predictive Value (NPV):** Measures out of everything classified as negative, how many are actually negative. High NPV is important to confidently rule out false negatives, ensuring that healthy fetuses are not mistakenly identified as at risk.
- **Specificity:** Measures how many actual negative cases were correctly identified. It ensures that negative cases are not falsely identified as positive.

Evaluation

Baseline Models

Logistic Regression (LR) is a linear model commonly used as a baseline in classification tasks due to its simplicity. LR demonstrated strong performance in identifying the majority 'normal' cases with high accuracy. However, its effectiveness diminished when applied to the minority classes ('suspect' and 'pathological'). While the AUC scores suggest that LR has good discriminative power, these results also underscore its limitations in dealing with class imbalances (see Figure 11).



Plain LR Accuracy Table

Class	Accuracy
normal	95.18%
suspect	75.00%
pathological	57.69%
Overall	89.34%

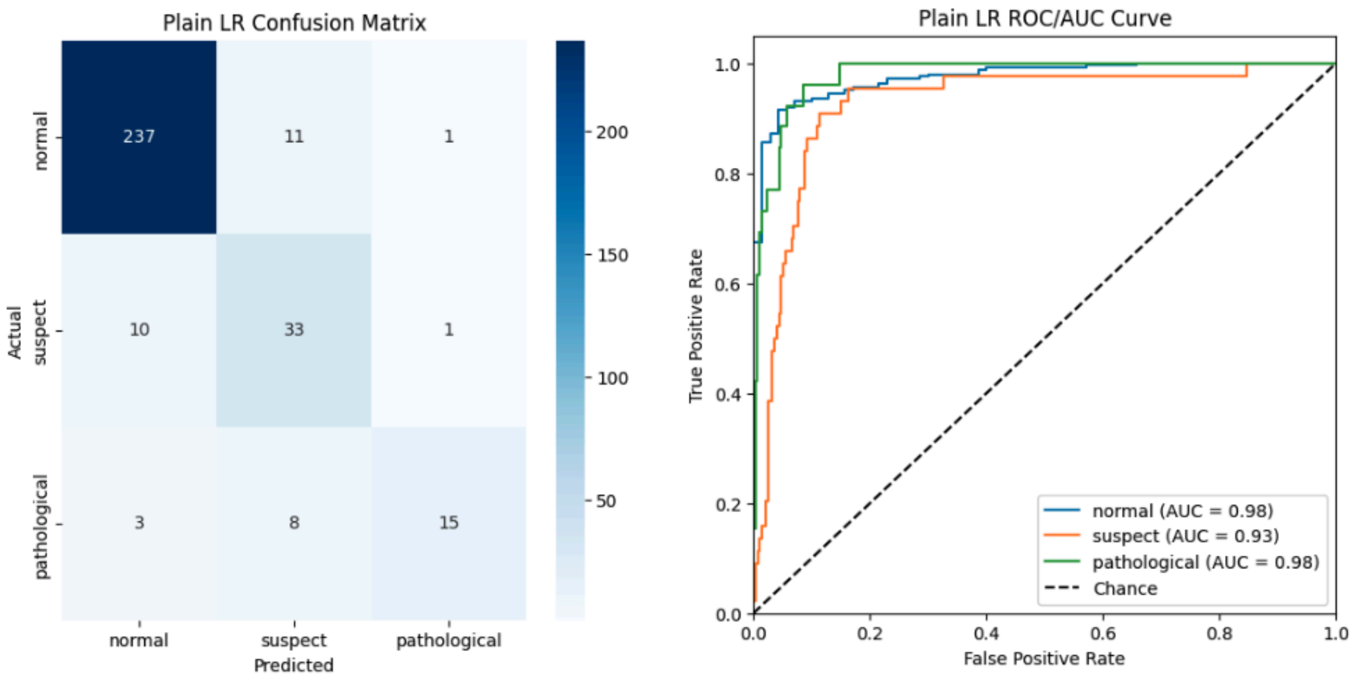
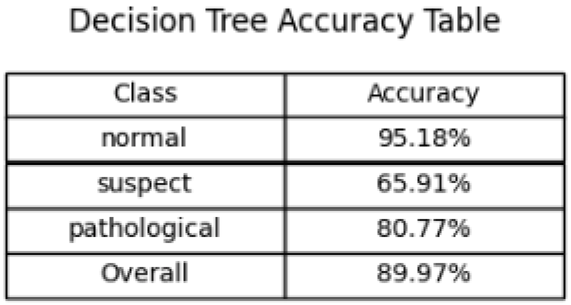


Figure 11: LR performance metrics: accuracy table, confusion matrix, and ROC/AUC curve

Decision Tree (DT) is a non-linear model that splits the dataset into branches to make predictions based on feature values to capture complex patterns in the data. DT displayed reasonable performance in predicting the 'normal' and 'pathological' classes. However, it struggled with the 'suspect' class. The ROC/AUC curves suggest that DTs discriminative power is not as strong as more sophisticated models (see Figure 12).



Decision Tree Accuracy Table

Class	Accuracy
normal	95.18%
suspect	65.91%
pathological	80.77%
Overall	89.97%

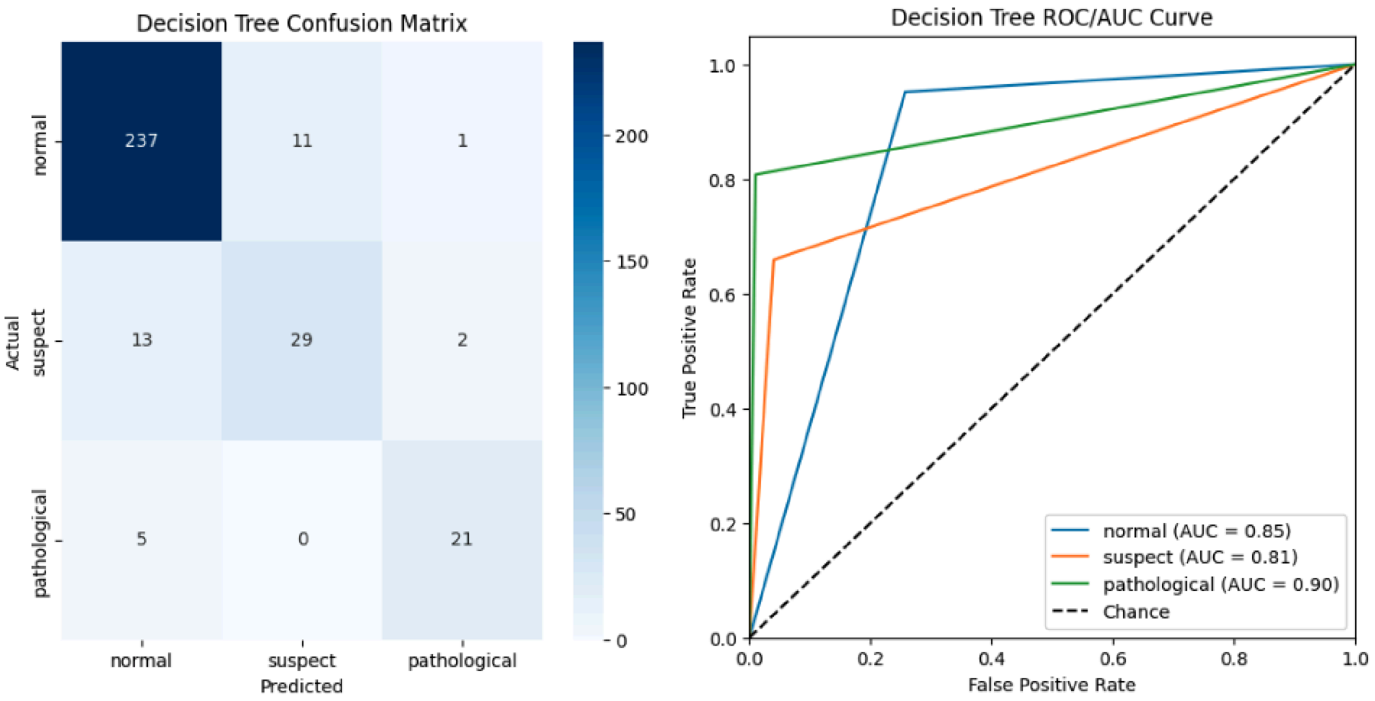
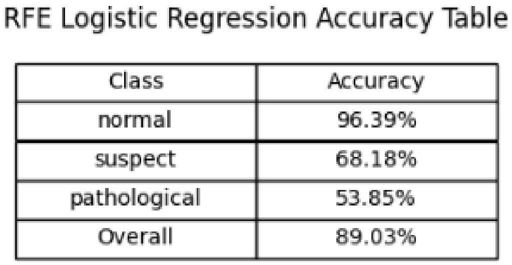


Figure 12: DT performance metrics: accuracy table, confusion matrix, and ROC/AUC curve.

Feature Selection

Recursive Feature Elimination (RFE) is a feature selection technique that works by recursively removing the least important features and building the model on those that remain. RFE selected 10 features: 'baseline value', 'accelerations', 'uterine contractions', 'prolonged decelerations', 'abnormal short-term variability', 'percentage of time with abnormal long-term variability', 'histogram mode', 'histogram mean', 'histogram median', and 'histogram variance'. These features are indicative of important aspects of the data that contribute to predicting fetal health. The performance of LR after applying RFE showed improved focus on the 'normal' class, with a slight trade-off in accuracy for the 'suspect' and 'pathological' classes. The ROC/AUC curves remain strong (see Figure 13).



RFE Logistic Regression Accuracy Table

Class	Accuracy
normal	96.39%
suspect	68.18%
pathological	53.85%
Overall	89.03%

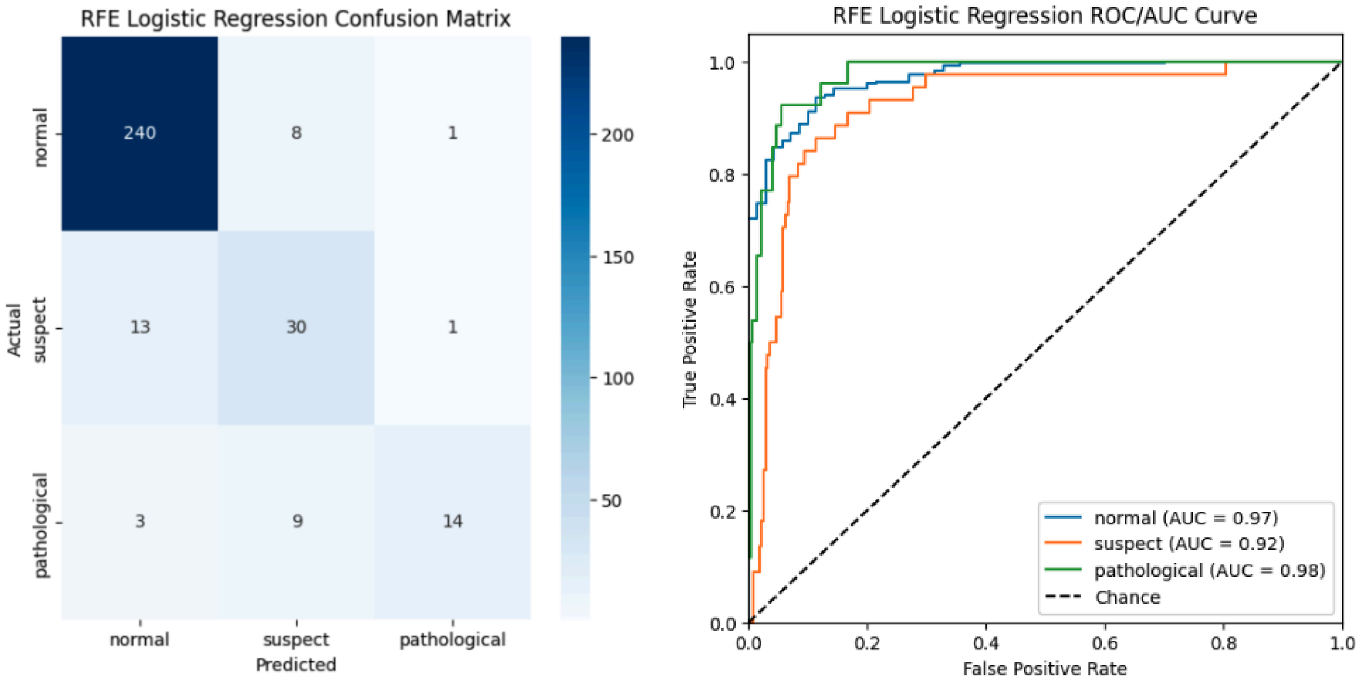
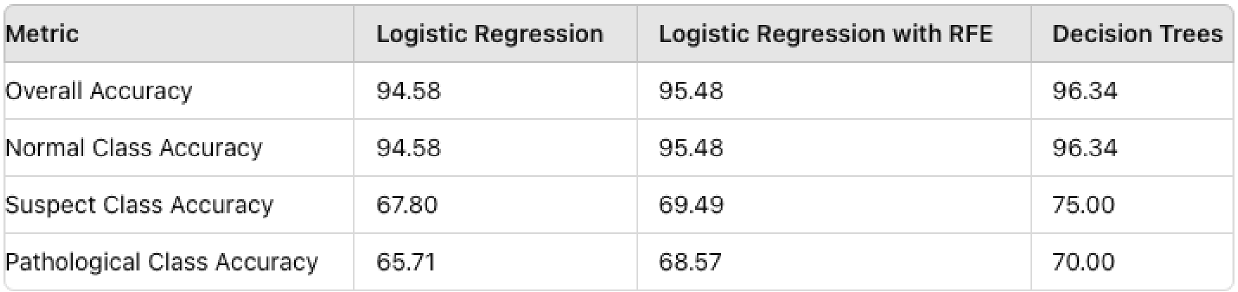


Figure 13: RFE LR performance metrics: accuracy table, confusion matrix, and ROC/AUC curve.

When comparing the accuracy metrics across the three models, LR, LR with RFE, and DT, several trends emerge. DT exhibited the highest overall accuracy and excelled in classifying the 'suspect' class. The introduction of RFE to LR resulted in an improvement in overall accuracy and in the classification of the 'suspect' and 'pathological' classes (see Figure 14).



Metric	Logistic Regression	Logistic Regression with RFE	Decision Trees
Overall Accuracy	94.58	95.48	96.34
Normal Class Accuracy	94.58	95.48	96.34
Suspect Class Accuracy	67.80	69.49	75.00
Pathological Class Accuracy	65.71	68.57	70.00

Figure 14: Comparison of accuracy metrics across LR, LR with RFE, and Decision Trees.

Class Imbalance

SMOTE (Synthetic Minority Over-sampling Technique) is used to generate synthetic samples for minority classes to address class imbalances in the dataset. SMOTE was applied to the training data with a sampling strategy that increased the minority class samples by 1.5%. The graph (see Figure 15) shows the feature distribution for Class 3 before and after resampling, indicating that the resampled data has a an evenly distributed feature set.

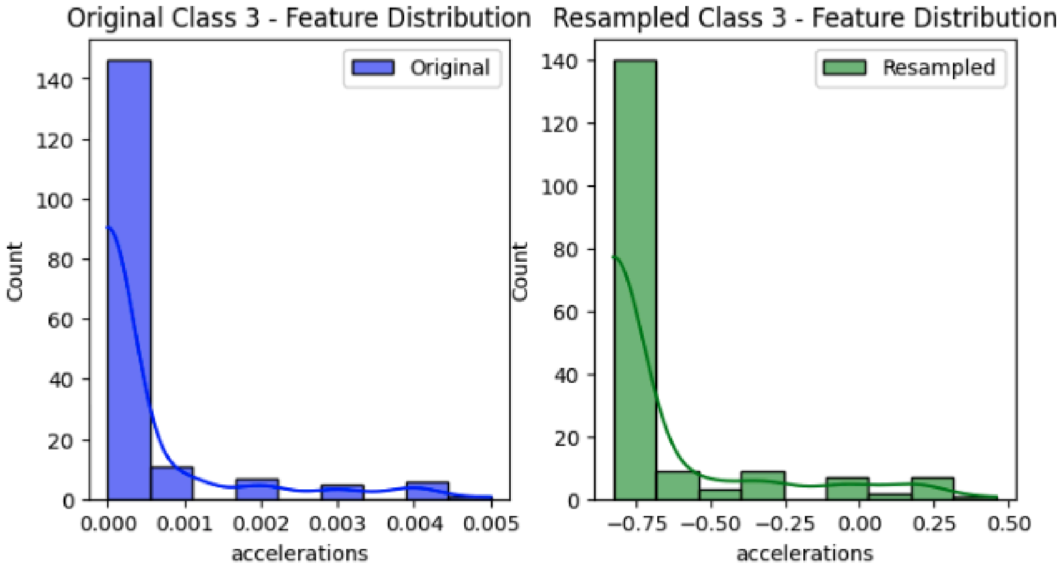
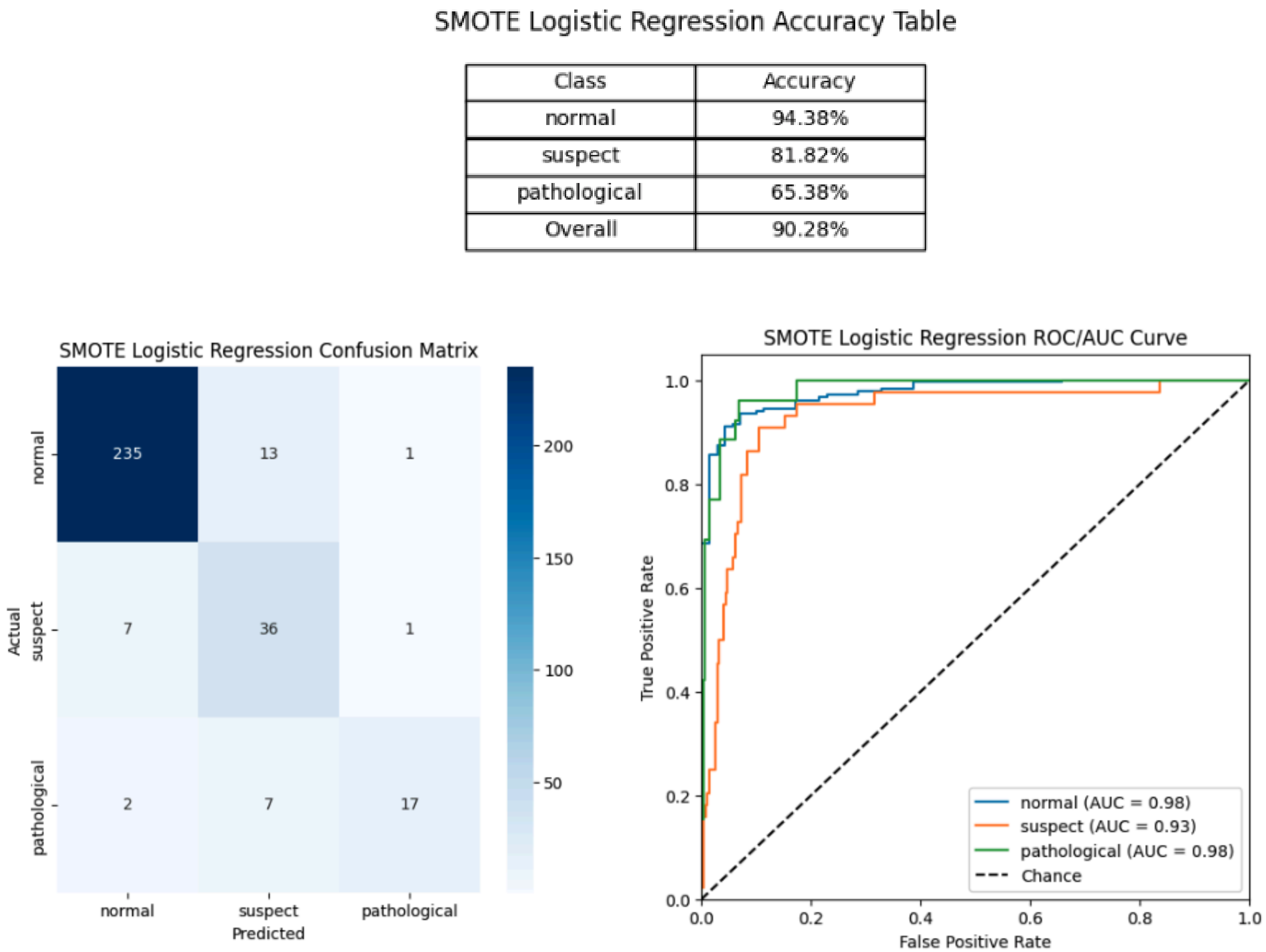


Figure 15: Feature distribution of Class 3 before and after SMOTE resampling.

The application of **SMOTE with LR** improved accuracy in predicting the 'suspect' and 'pathological' classes, and the overall accuracy. The ROC/AUC curves indicate that the model's ability to distinguish between classes remained strong (see Figure 16).



SMOTE Logistic Regression Accuracy Table

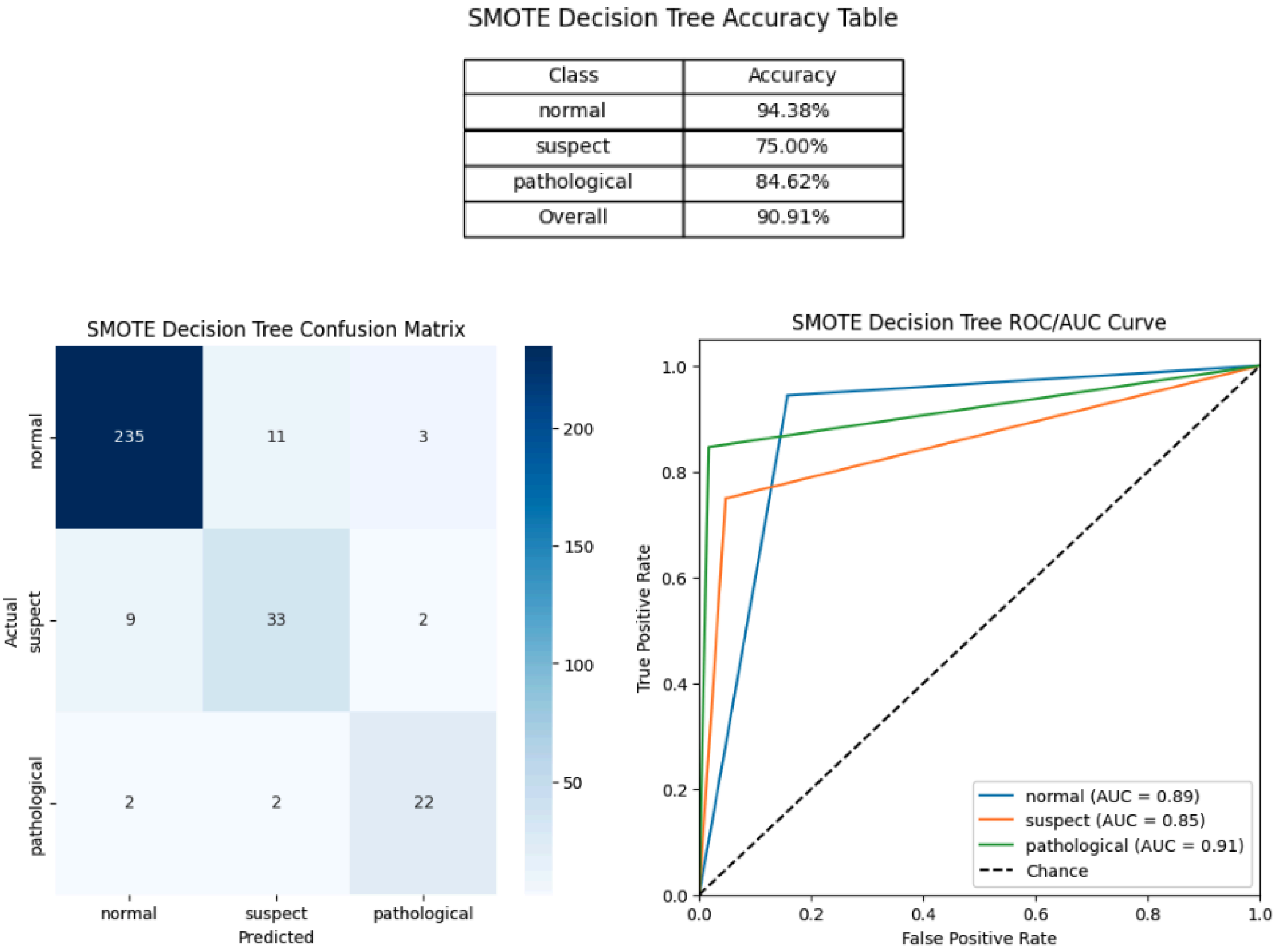
Class	Accuracy
normal	94.38%
suspect	81.82%
pathological	65.38%
Overall	90.28%

SMOTE Logistic Regression Confusion Matrix

Actual \ Predicted	normal	suspect	pathological
normal	235	13	1
suspect	7	36	1
pathological	2	7	17

Figure 16: SMOTE LR performance metrics: accuracy table, confusion matrix, and ROC/AUC curve.

DT with SMOTE showed significant improvement in handling the minority classes. The model achieved a balanced overall accuracy, with particularly strong performance in the 'normal' and 'pathological' classes (see Figure 17).

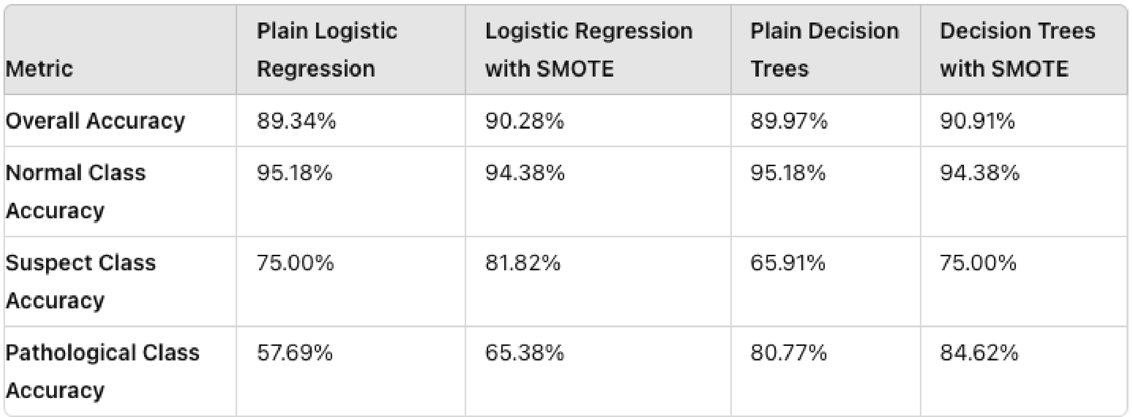


SMOTE Decision Tree Accuracy Table

Class	Accuracy
normal	94.38%
suspect	75.00%
pathological	84.62%
Overall	90.91%

Figure 17: SMOTE DT performance metrics: accuracy table, confusion matrix, and ROC/AUC curve.

The comparison between plain LR and DT models versus their SMOTE-enhanced versions shows that applying SMOTE generally improves the models' performance on minority classes. However, SMOTE was not used for this medical dataset. Synthetic data generation, especially in medical data, can introduce biases or overfitting. Even though SMOTE enhanced performance in some areas, the potential risks associated with synthetic data in this sensitive field outweighed the benefits (see Figure 18).

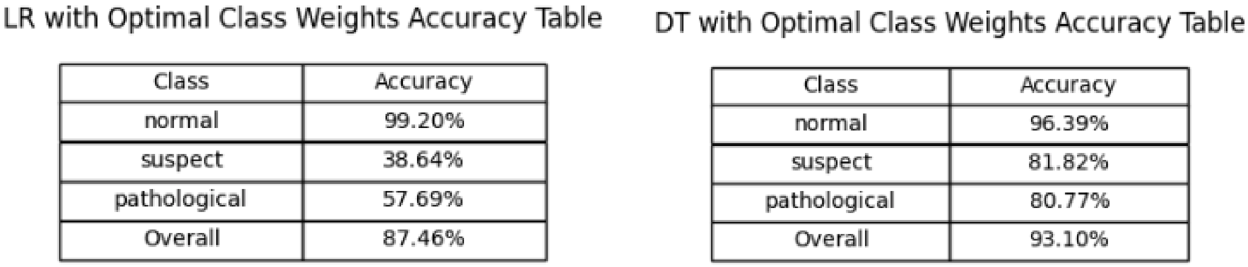


Metric	Plain Logistic Regression	Logistic Regression with SMOTE	Plain Decision Trees	Decision Trees with SMOTE
Overall Accuracy	89.34%	90.28%	89.97%	90.91%
Normal Class Accuracy	95.18%	94.38%	95.18%	94.38%
Suspect Class Accuracy	75.00%	81.82%	65.91%	75.00%
Pathological Class Accuracy	57.69%	65.38%	80.77%	84.62%

Figure 18: Comparison of performance metrics between plain and SMOTE-enhanced models.

Weighted Penalty

A **weighted penalty approach** was implemented for both Logistic Regression and Decision Trees. For LR, this approach significantly boosted the model's ability to classify the 'normal' class with high precision but significantly decreased the performance for the 'suspect' class. For DT, this approach showed enhanced performance across all classes. It was particularly effective in improving the detection of the 'suspect' and 'pathological' classes, demonstrating better sensitivity and specificity. The ROC/AUC curves reflected a balanced and strong discriminative ability (see Figure 19).



LR with Optimal Class Weights Accuracy Table

Class	Accuracy
normal	99.20%
suspect	38.64%
pathological	57.69%
Overall	87.46%

DT with Optimal Class Weights Accuracy Table

Class	Accuracy
normal	96.39%
suspect	81.82%
pathological	80.77%
Overall	93.10%

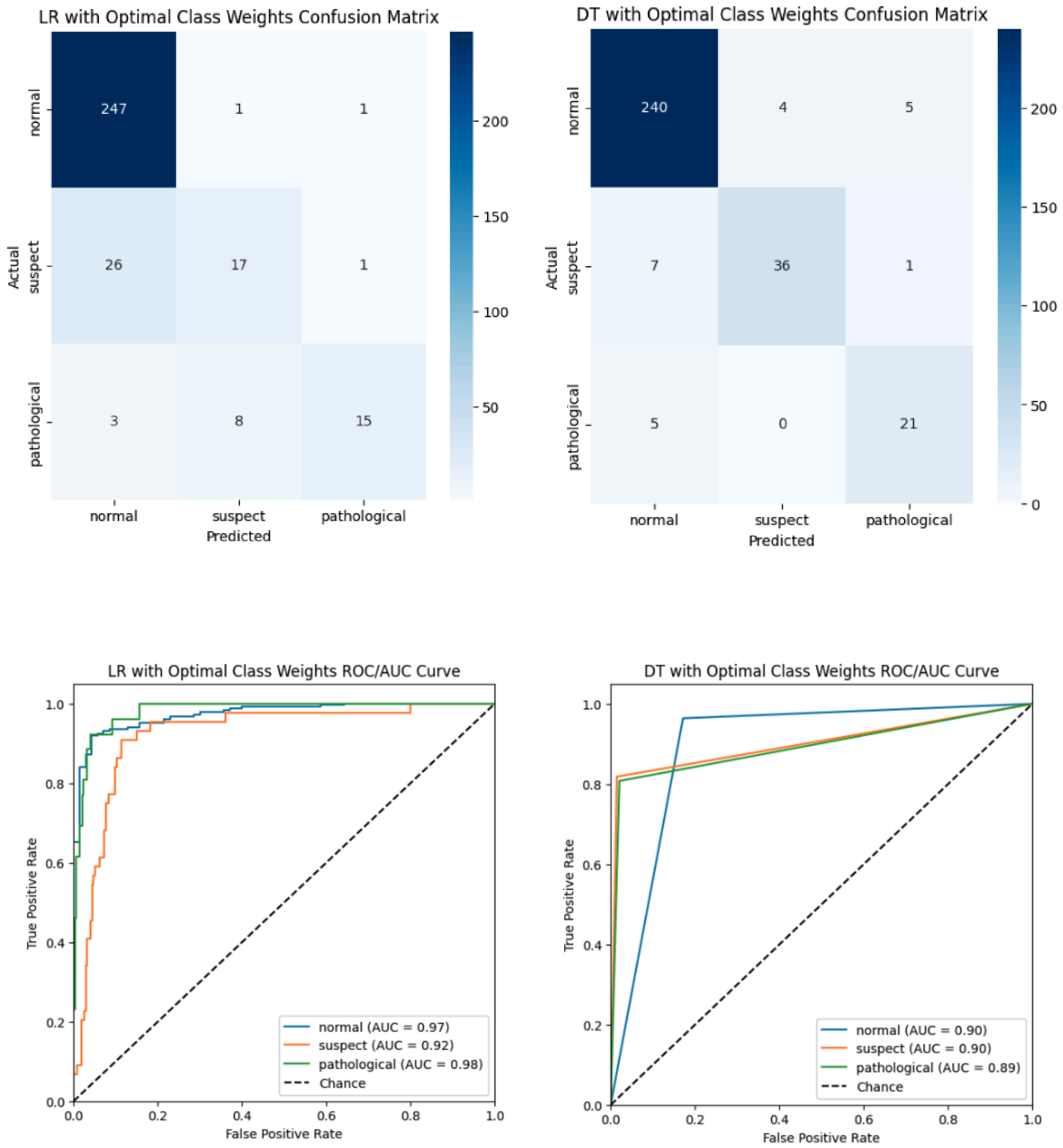
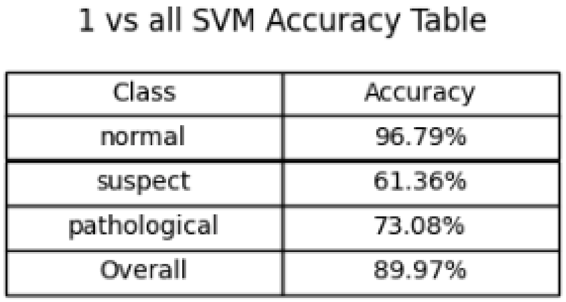


Figure 19: Weighted penalty metrics for LR and DT with class weights.

One-vs-All SVM Approach

One-vs-All Support Vector Machine (SVM) is an approach used for multi-class classification where multiple binary SVM models are trained (one for each class). Each model separates one class from the rest, and during prediction, the class with the highest probability is chosen as the final prediction. This involved training three separate SVM models, each distinguishing one of the three classes (normal, suspect, and pathological) from the others. The results indicate that the 1-vs-all SVM model achieved strong overall accuracy, particularly excelling in classifying the 'normal' class. The ROC/AUC curves demonstrate good discriminative power across all classes (see Figure 20).



1 vs all SVM Accuracy Table

Class	Accuracy
normal	96.79%
suspect	61.36%
pathological	73.08%
Overall	89.97%

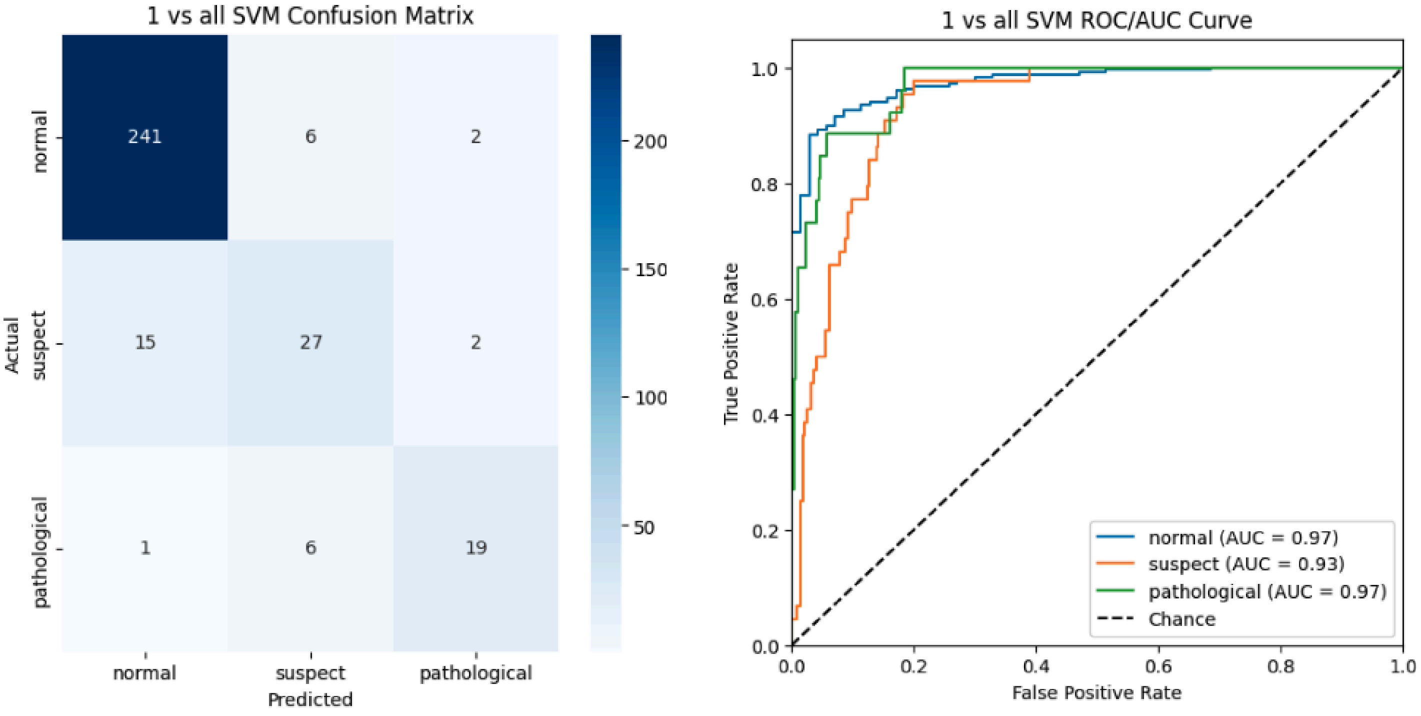
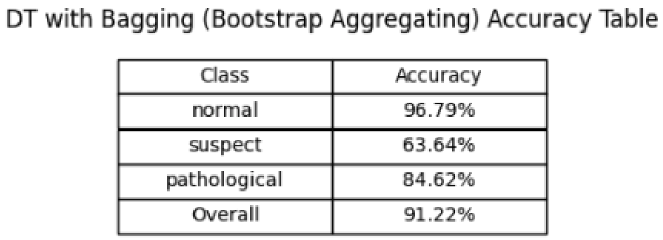


Figure 20: 1-vs-all SVM performance metrics.

Decision Trees: Bagging Techniques

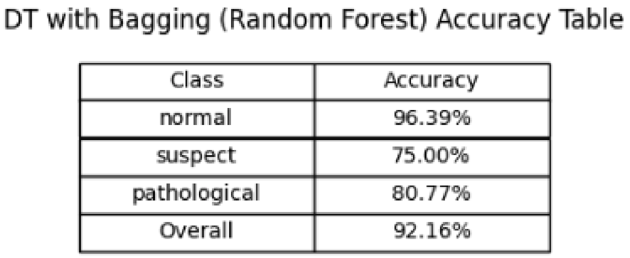
Bagging improves model stability and accuracy by training multiple versions of the same model on different subsets of the data and then combining their predictions. In **Bootstrap Aggregating**, multiple Decision Trees are trained independently on various subsets of the training data, created through bootstrapping (sampling with replacement). The predictions from these individual trees are then aggregated to produce the final output. This approach performed well, particularly in classifying the 'normal' and 'pathological' classes, as evidenced by strong ROC/AUC curves across all classes.

Random Forest introduces an additional layer of randomness: at each split in a tree, only a random subset of features is considered. It handles the 'suspect' class more effectively. The inclusion of feature randomization helped the model achieve higher discriminative power, as reflected in the ROC/AUC curves, and slightly better overall accuracy compared to Bagging with Decision Trees (see Figure 21).



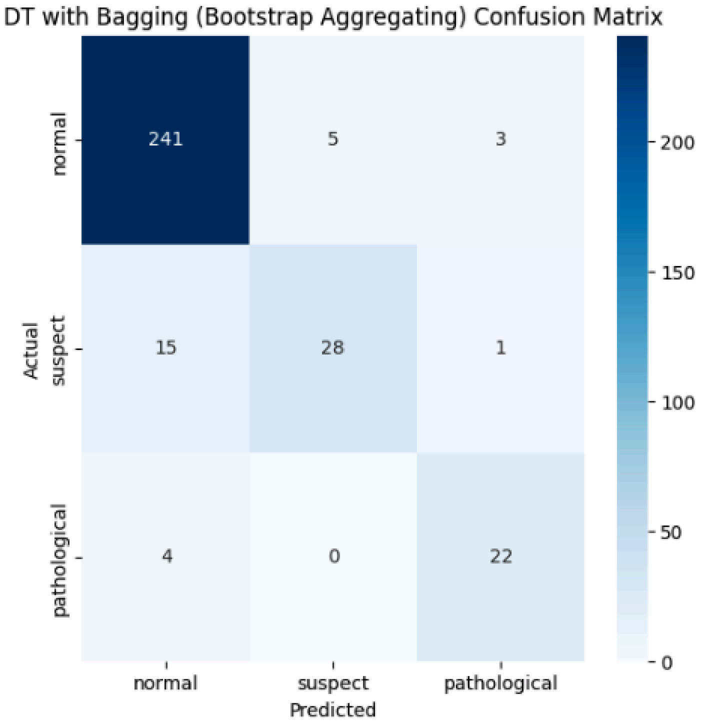
DT with Bagging (Bootstrap Aggregating) Accuracy Table

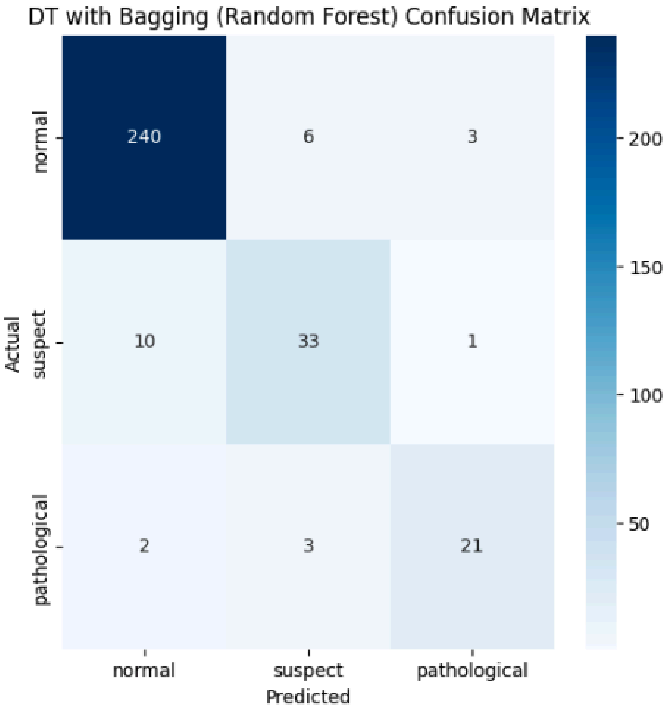
Class	Accuracy
normal	96.79%
suspect	63.64%
pathological	84.62%
Overall	91.22%



DT with Bagging (Random Forest) Accuracy Table

Class	Accuracy
normal	96.39%
suspect	75.00%
pathological	80.77%
Overall	92.16%





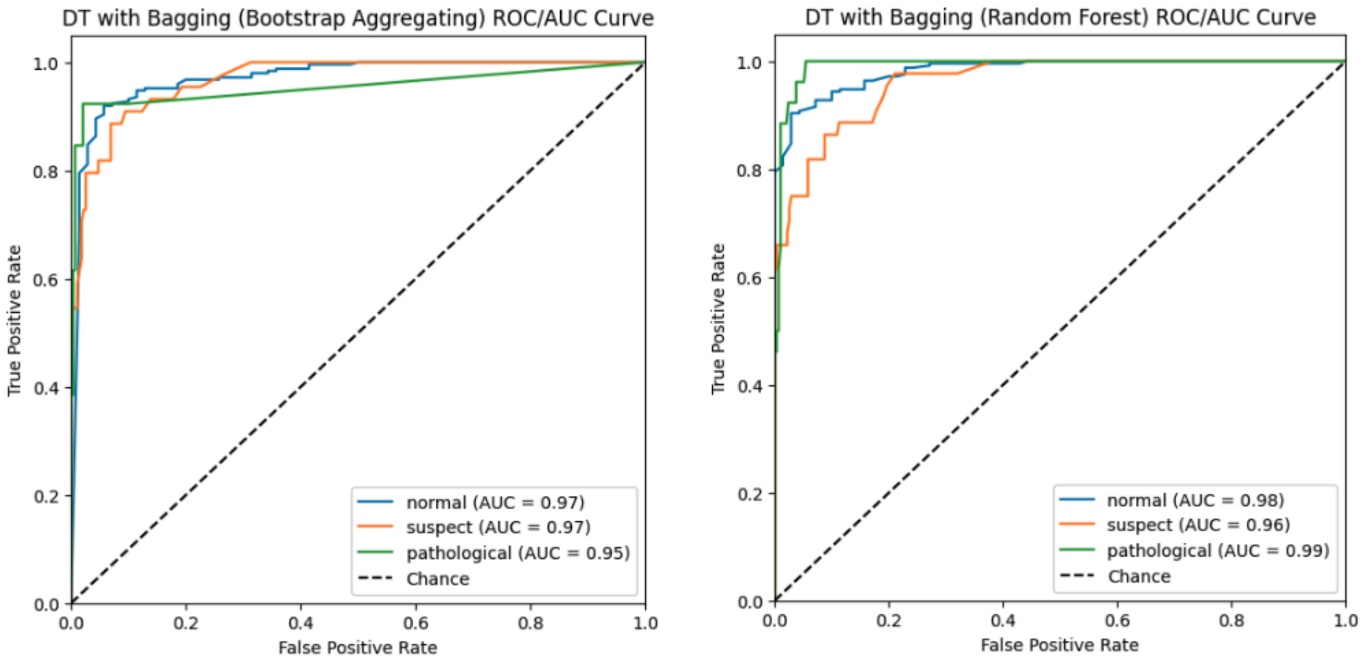
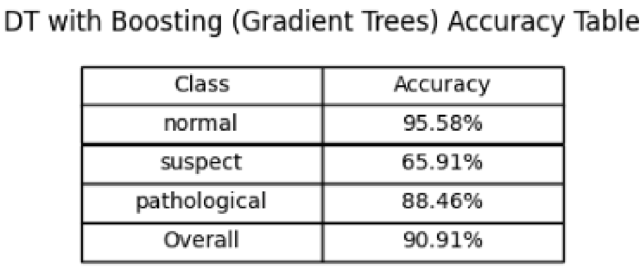


Figure 21: Bagging vs. Random Forest performance metrics.

Decision Trees: Boosting Techniques

Gradient Boosting works by building decision trees sequentially, where each subsequent tree aims to correct the errors made by the previous ones. This model demonstrated strong performance in classifying the 'pathological' class but not with the 'suspect' class. The ROC/AUC curves confirm that the model has good discriminative power across all classes (see Figure 22).



DT with Boosting (Gradient Trees) Accuracy Table

Class	Accuracy
normal	95.58%
suspect	65.91%
pathological	88.46%
Overall	90.91%

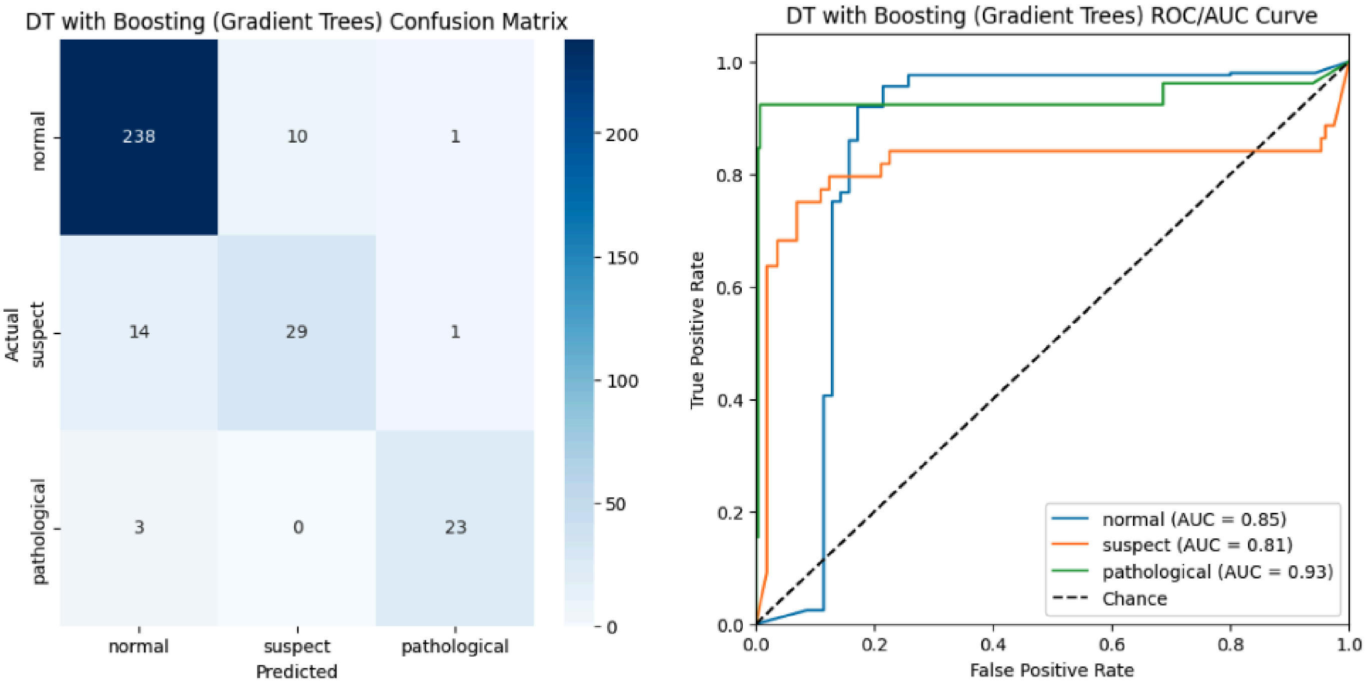
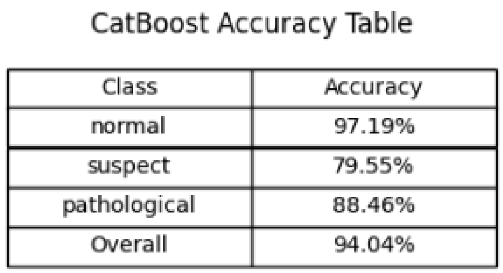


Figure 22: Gradient Boosting performance metrics.

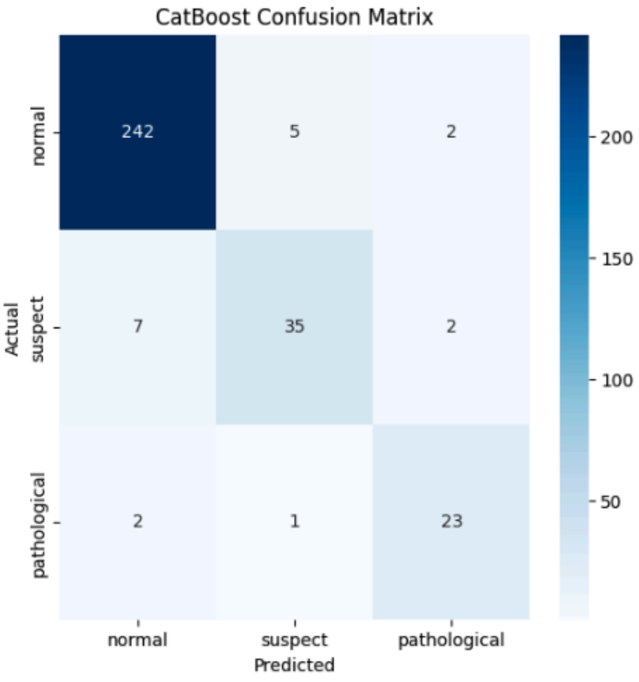
CatBoost Classifier

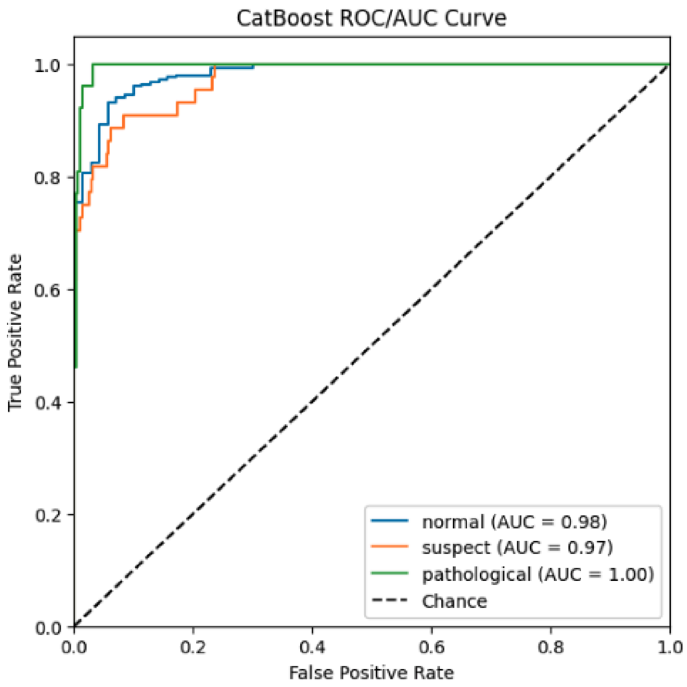
CatBoost specializes in handling categorical data, using gradient boosting with symmetric trees. This model demonstrated strong overall performance, particularly excelling in the 'normal' and 'pathological' classes, as shown by the high accuracy and ROC/AUC scores across these classes. The model achieved an overall accuracy of 94.04%, reaching a perfect AUC of 1.00 for the 'pathological' class. To further optimize the model, different class weight settings were experimented with using grid search to adjust the model's sensitivity to imbalanced classes. However, this approach did not yield any significant improvements in performance, indicating that the base model's handling of class imbalance was already sufficient for this dataset (see Figure 23).



CatBoost Accuracy Table

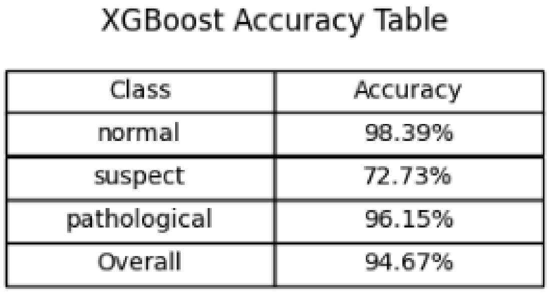
Class	Accuracy
normal	97.19%
suspect	79.55%
pathological	88.46%
Overall	94.04%





XGBoost Classifier

XGBoost is another robust gradient boosting algorithm that works by training multiple decision trees sequentially. The results demonstrated strong performance, with an overall accuracy of 94.67%. The model particularly excelled in classifying the 'normal' and 'pathological' classes, achieving high ROC/AUC scores of 0.98 and 0.99, respectively. To further optimize the model, different class weight settings were experimented with using grid search to adjust the model's sensitivity to imbalanced classes. However, this approach did not yield any significant improvements in performance, indicating that the base model's handling of class imbalance was already sufficient for this dataset (see Figure 24).



XGBoost Accuracy Table

Class	Accuracy
normal	98.39%
suspect	72.73%
pathological	96.15%
Overall	94.67%

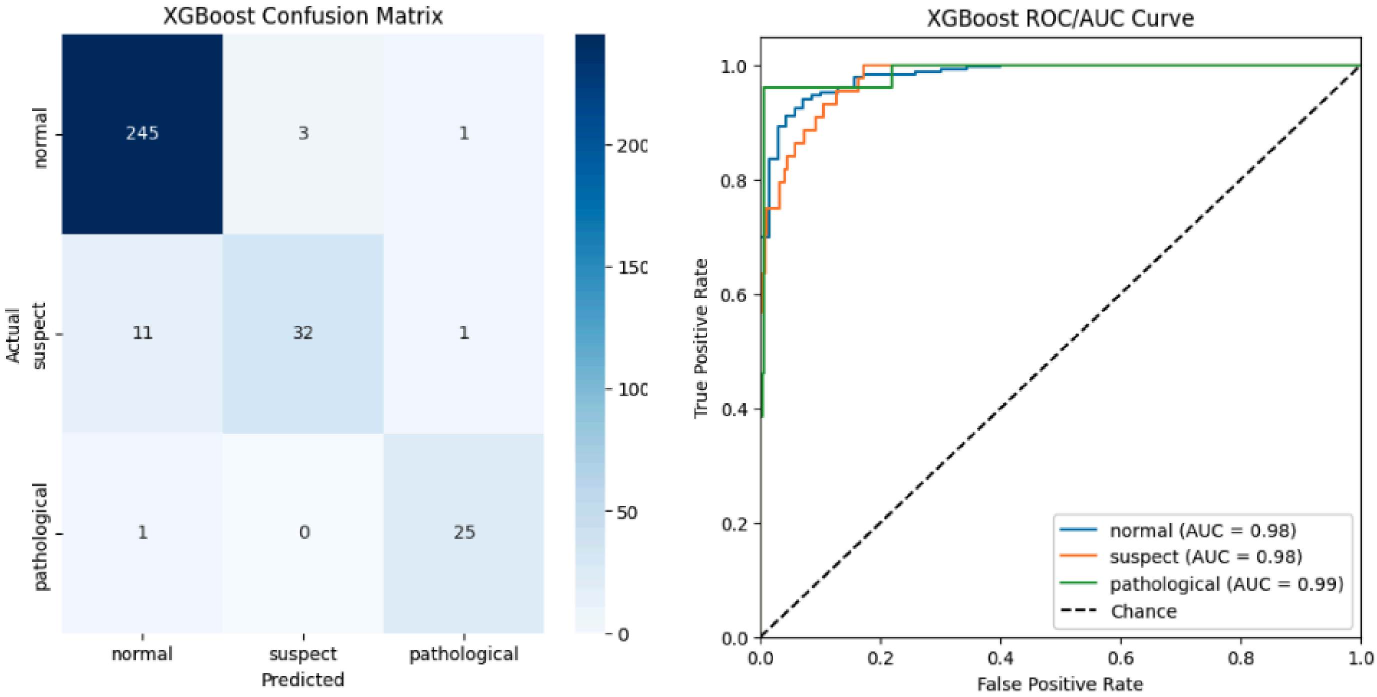
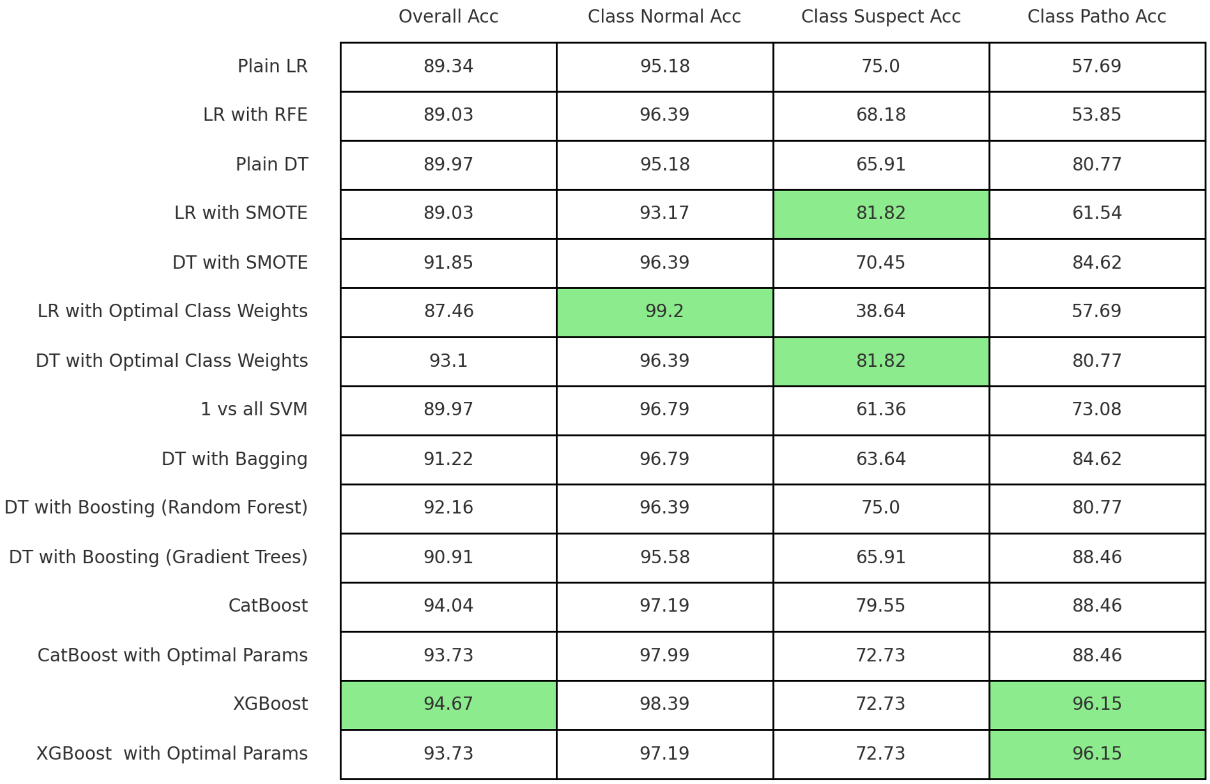


Figure 24: XGBoost performance metrics.

Ensemble Model

Rationale

The baseline evaluation highlighted the strengths and limitations of various models. While LR and DT served as solid baselines, they struggled with minority classes. Feature selection with RFE improved LR's performance in the 'normal' class, but linear models still showed limitations with imbalanced datasets. SMOTE improved class balance handling but was not used in the final models due to concerns about synthetic data in medical contexts. Weighted penalties enhanced both LR and DT. Random Forest provided better handling of the 'suspect' class and reduced overfitting by training multiple decision trees on bootstrapped subsets. Advanced models like CatBoost and XGBoost exhibited strong overall performance, with XGBoost emerging as the model with the highest overall accuracy (see Figure 25).



	Overall Acc	Class Normal Acc	Class Suspect Acc	Class Patho Acc
Plain LR	89.34	95.18	75.0	57.69
LR with RFE	89.03	96.39	68.18	53.85
Plain DT	89.97	95.18	65.91	80.77
LR with SMOTE	89.03	93.17	81.82	61.54
DT with SMOTE	91.85	96.39	70.45	84.62
LR with Optimal Class Weights	87.46	99.2	38.64	57.69
DT with Optimal Class Weights	93.1	96.39	81.82	80.77
1 vs all SVM	89.97	96.79	61.36	73.08
DT with Bagging	91.22	96.79	63.64	84.62
DT with Boosting (Random Forest)	92.16	96.39	75.0	80.77
DT with Boosting (Gradient Trees)	90.91	95.58	65.91	88.46
CatBoost	94.04	97.19	79.55	88.46
CatBoost with Optimal Params	93.73	97.99	72.73	88.46
XGBoost	94.67	98.39	72.73	96.15
XGBoost with Optimal Params	93.73	97.19	72.73	96.15

Figure 25: Comparison of overall and class-specific accuracies across models.

Logistic Regression (LR) with Optimal Class Weights was the best model for the Normal class. Its ability to maintain a balanced precision ensures the most significant features were accurately weighted and that the normal cases were correctly classified without being overshadowed. **Decision Tree (DT) with Optimal Class Weights** was the best model for the Suspect class, offering better interpretability and handling of imbalanced data by adjusting weights for minority classes. **XGBoost** demonstrated superior performance for the Pathological class, achieving the highest overall accuracy across all categories while maintaining efficiency in training time compared to models like CatBoost.

Ensemble models can be highly beneficial for complex tasks like fetal health classification, where data imbalance and varying feature importance across classes present significant challenges. While Random Forest handles feature importance effectively, combining models in an ensemble allows for better overall accuracy and robustness across all classes. An ensemble approach allows leveraging the strengths of multiple models, each tailored to perform in a specific class. Both ensemble methods were trained on the same data split to maintain the integrity of class distributions.

Ensemble Data Preprocessing

Data handling followed a structure to ensure objective evaluation and prevent model leakage. All individual models in the ensemble were given training and validation sets derived from the same original data split. These models were trained and tuned independently before being integrated into the ensemble. Importantly, each ensemble was evaluated on a completely separate test set that none of the individual submodels had been exposed to during training or validation. This universal test set was essential to ensure that model evaluation was fair and unbiased across all submodels.

The dataset was initially split with 70% (1,488 samples) allocated for training. The remaining 30% (638 samples) was further split equally into 15% (319 samples) for validation and 15% (319 samples) for testing. This approach ensured that the test set was exclusively reserved for final evaluation, while the training data was split into training and validation subsets to optimize hyperparameters during model development. Both ensemble models used this consistent split, with stratified sampling to maintain the class distribution across all subsets. The universal test set (319 samples) was maintained throughout the process, ensuring that no model was exposed to it during training or validation. This provided a single, standardized benchmark for objectively comparing model performance across different approaches.

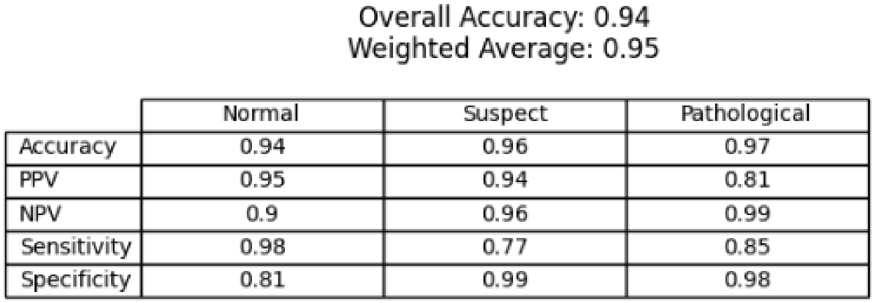
Ensemble Prioritized by Overall Accuracy

This method incorporates a hierarchical decision-making process:

- XGBoost takes precedence if its prediction confidence exceeds a predefined high-confidence threshold (95%). This model was given priority due to its robust performance. If XGBoost's confidence is below the threshold, the decision defaults to the other models.
- Decision Tree (DT) with optimal class weights is used if it shows high confidence, especially in identifying the "Suspect" class.
- Logistic Regression (LR) with optimal class weights is utilized if it exhibits high confidence, particularly in the "Normal" class.

In cases where the models disagree and no single model's confidence is high enough, the ensemble prioritizes XGBoost's prediction, particularly if it predicts the "Pathological" class. If XGBoost is uncertain, the ensemble defers to the Decision Tree for "Suspect" class predictions or Logistic Regression for "Normal" class predictions. This strategy ensures that the most reliable model influences the final decision.

This ensemble models overall accuracy and weighted average accuracy are high and the confusion matrix shows that the model excels particularly in classifying the "Normal" and "Pathological" classes, with only a small number of misclassifications. However, the "Suspect" class shows some challenges, with a lower sensitivity compared to the other classes. This indicates that the model occasionally struggles to differentiate "Suspect" cases from "Normal" and "Pathological". This could be attributed to the overlapping features between "Suspect" and the other two classes, which makes it more difficult for the model to draw clear distinctions. Despite this, the model's performance for the "Suspect" class remains solid, as evidenced by the ROC/AUC score. The ROC/AUC curves highlight the model's discriminative power, with solid curves for the "Normal" and "Pathological" classes (see Figure 26).



Overall Accuracy: 0.94
Weighted Average: 0.95

	Normal	Suspect	Pathological
Accuracy	0.94	0.96	0.97
PPV	0.95	0.94	0.81
NPV	0.9	0.96	0.99
Sensitivity	0.98	0.77	0.85
Specificity	0.81	0.99	0.98

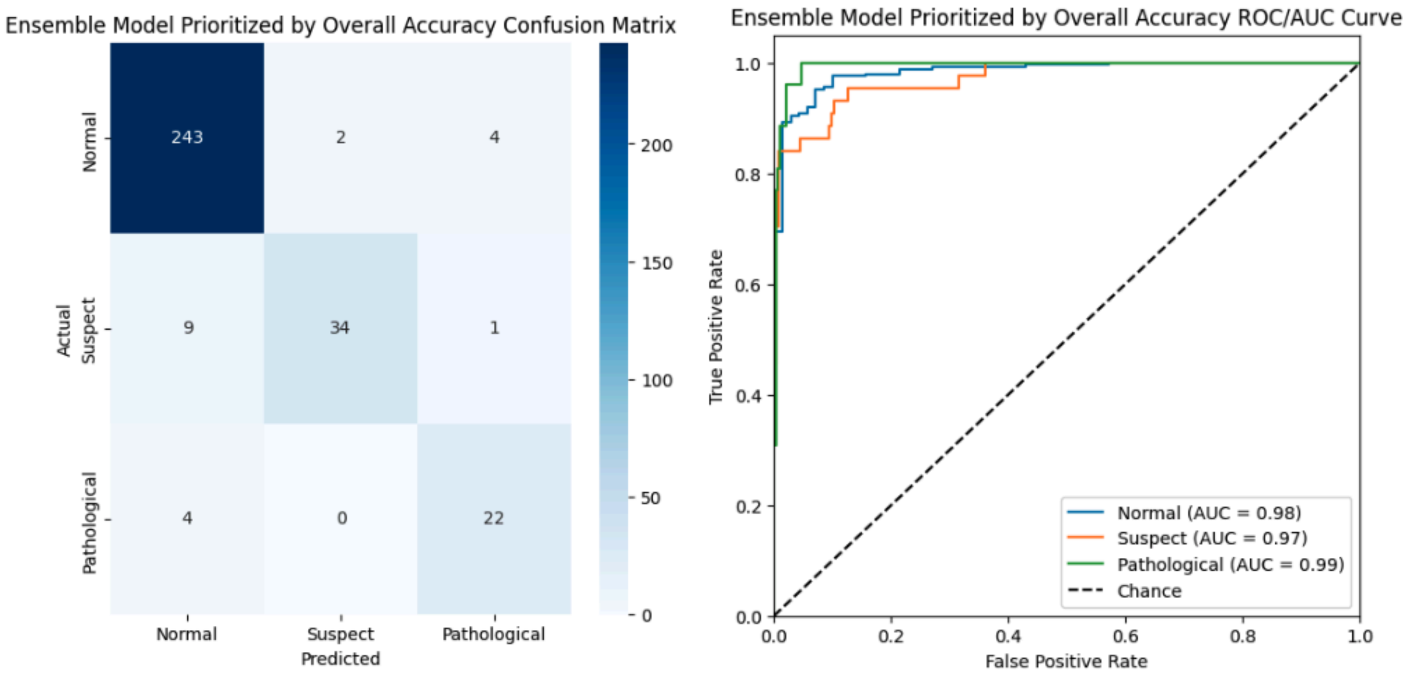


Figure 26: Ensemble model performance metrics.

Upon examining the misclassified instances from this model, the majority of misclassifications involve the "Suspect" class, with 11 out of the 20 misclassified instances belonging to this category. The model struggles to differentiate between the "Suspect" and the other two classes, particularly "Normal." A common characteristic among the misclassified "Suspect" instances is the presence of borderline feature values, such as slightly elevated or reduced baseline FHR and minimal decelerations or accelerations. Additionally, several "Normal" instances were incorrectly classified as "Pathological," suggesting that extreme feature values, such as lower baseline FHR combined with the absence of fetal movement or accelerations, led to a more severe health status.

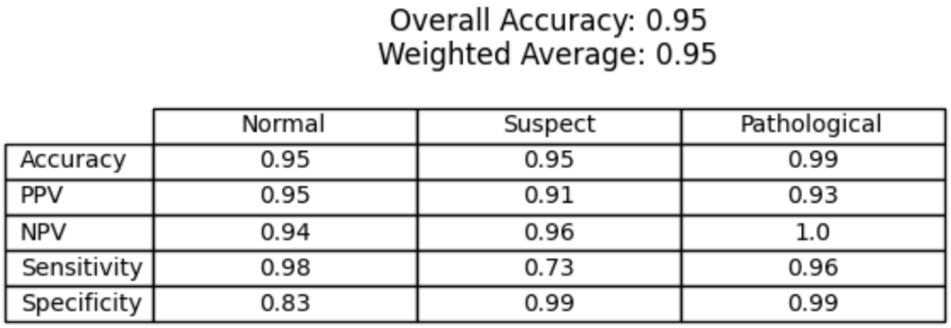
Voting Ensemble Model

This method implements a weighted voting mechanism to determine the final prediction:

- During the prediction phase, each model casts a "vote" for one of the three possible classes. XGBoost is given an additional vote due to its superior overall accuracy.
- The Decision Tree and Logistic Regression models each contribute one vote.
- The final prediction is determined by the class that receives the most votes. If there is a tie or if the models predict different classes, the class favored by XGBoost becomes the final prediction.

This method ensures that the ensemble benefits from the collective models while giving extra weight to the most accurate model, XGBoost, during the decision-making process.

This ensemble model shows stronger overall performance, with the confusion matrix revealing that the model effectively distinguishes between the 'Normal' and 'Pathological' classes. However, the 'Suspect' class presents a challenge. The ROC/AUC curve shows the model performs well in distinguishing the 'Normal' and 'Pathological' classes, but the curve for the 'Suspect' class is lower. While the 'Normal' and 'Pathological' classes maintain high performance across all metrics, the 'Suspect' class shows a drop in sensitivity (see Figure 27).



Overall Accuracy: 0.95
Weighted Average: 0.95

	Normal	Suspect	Pathological
Accuracy	0.95	0.95	0.99
PPV	0.95	0.91	0.93
NPV	0.94	0.96	1.0
Sensitivity	0.98	0.73	0.96
Specificity	0.83	0.99	0.99

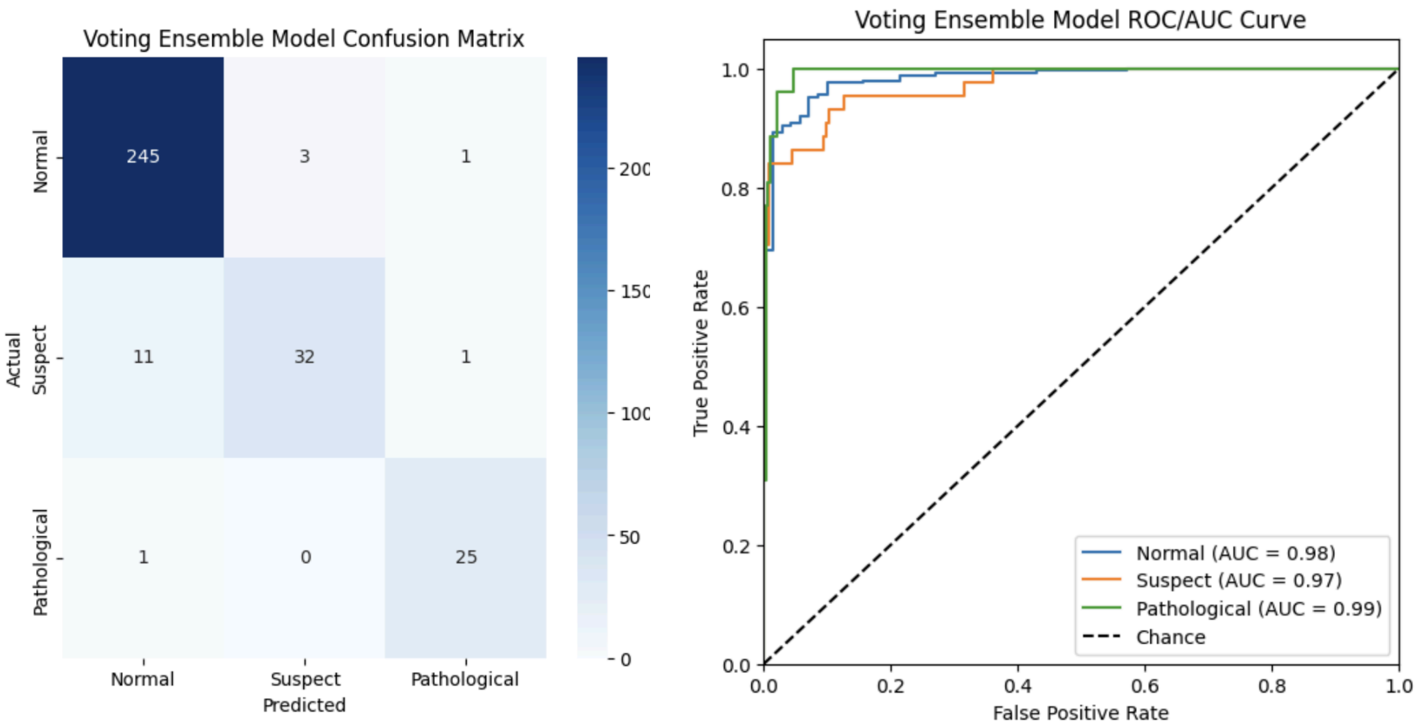


Figure 27: Voting ensemble model performance metrics.

The misclassifications observed in this model highlight challenges primarily associated with distinguishing between the "Suspect" and "Normal" classes. Among the 17 misclassified instances, a significant number involved the "Suspect" class being incorrectly predicted as "Normal." This suggests that the model struggles to detect subtle variations in features that distinguish "Suspect" cases from "Normal" ones, particularly when these cases exhibit borderline feature values. There are instances where "Normal" cases were misclassified as "Suspect”. There was also a case where a "Normal" instance was misclassified as "Pathological”. These misclassifications suggest that the model faces specific challenges in accurately classifying cases within the "Suspect" category.

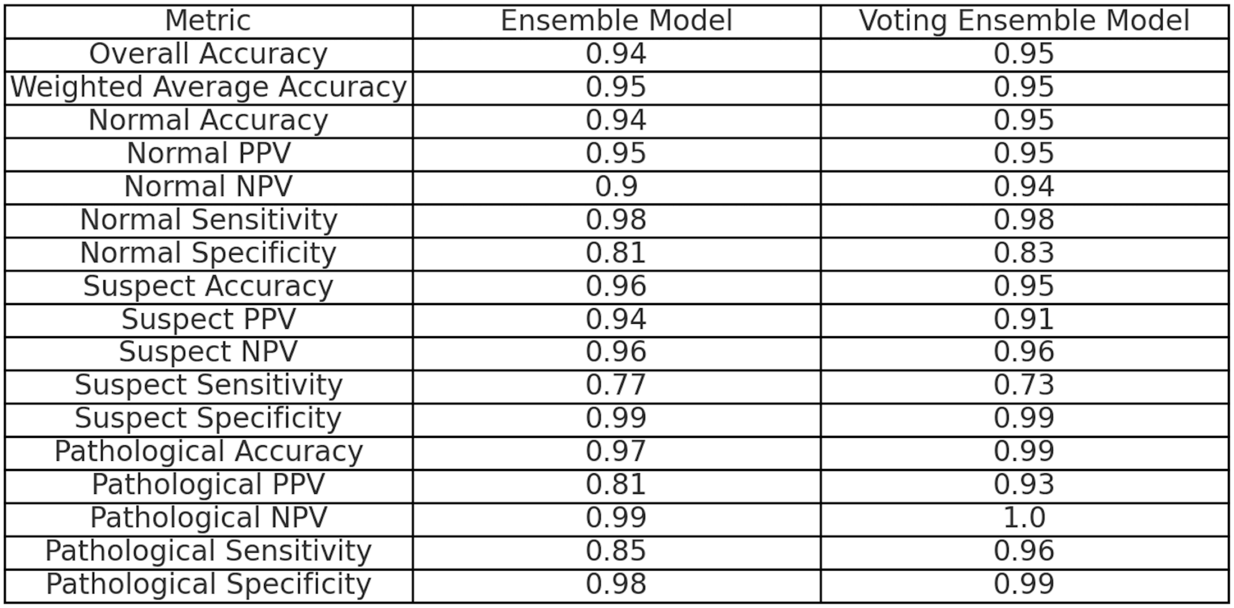
Comparison

Both ensemble models show similar performance in overall accuracy and weighted average accuracy, with the Voting Ensemble slightly outperforming the Accuracy Ensemble (0.95 vs. 0.94). Despite this, both models maintain the same weighted average accuracy (see Figure 28).

For the Normal class, the Voting Ensemble has a slight advantage in accuracy, NPV, and specificity - making it more reliable in minimizing false negatives. Both models share identical PPV and sensitivity, reflecting comparable performance in flagging normal instances.

For the Suspect class, the Ensemble Model achieves higher sensitivity (0.77 vs. 0.73), indicating better detection of suspect cases. The Voting Ensemble shows slightly lower PPV (0.91 vs. 0.94), meaning it may misclassify more suspect cases. Both models struggle with sensitivity in this class, revealing a shared weakness.

The Voting Ensemble excels in the Pathological class, with improved accuracy, sensitivity, and PPV - highlighting its strength in minimizing both false negatives and false positives. The Ensemble Model shows lower sensitivity and PPV - making it less robust in detecting pathological cases.



Metric	Ensemble Model	Voting Ensemble Model
Overall Accuracy	0.94	0.95
Weighted Average Accuracy	0.95	0.95
Normal Accuracy	0.94	0.95
Normal PPV	0.95	0.95
Normal NPV	0.9	0.94
Normal Sensitivity	0.98	0.98
Normal Specificity	0.81	0.83
Suspect Accuracy	0.96	0.95
Suspect PPV	0.94	0.91
Suspect NPV	0.96	0.96
Suspect Sensitivity	0.77	0.73
Suspect Specificity	0.99	0.99
Pathological Accuracy	0.97	0.99
Pathological PPV	0.81	0.93
Pathological NPV	0.99	1.0
Pathological Sensitivity	0.85	0.96
Pathological Specificity	0.98	0.99

Figure 28: Comparison of ensemble model and voting ensemble model metrics.

Conclusion

This research focused on the classification of fetal health status using CTG signals, employing a range of machine learning models. The Voting Ensemble model demonstrated superior performance, achieving high accuracy in identifying the "Normal" and "Pathological" classes. The study highlights the challenges associated with accurately classifying the "Suspect" class.

The Voting Ensemble Model is the best choice for fetal health classification because it offers a balanced and robust performance across all classes, particularly excelling in the "Pathological" and "Normal" categories. It achieves higher overall accuracy and demonstrates superior sensitivity and PPV for the "Pathological" class, making it highly effective in identifying cases where fetal health is at serious risk. This model also maintains a strong performance in the "Normal" class, minimizing both false positives and false negatives. While the "Suspect" class poses challenges for all models, the Voting Ensemble still manages to provide competitive results. This balanced and comprehensive approach makes the Voting Ensemble Model the most suitable choice for improving prenatal care.

In clinical practice, it is often preferable to have false positives rather than false negatives. It is preferable to inform a pregnant woman that there may be a potential issue, even if it ultimately proves to be false, rather than assuring her that everything is fine when, in fact, there is a problem. The Voting Ensemble Model aligns well with this preference, as it is designed to minimize false negatives, particularly in critical "Pathological" cases.

The implementation of these models has the potential to significantly enhance prenatal care. By accurately classifying fetal health status, these models can help healthcare providers identify cases of fetal distress more effectively, reducing unnecessary medical interventions and improving pregnancy outcomes. The use of such models in clinical settings could contribute to lowering neonatal and maternal mortality rates, aligning with global health objectives.

References

1. Faundes, A., Menezes, G., & Osis, M. J. (2019). Corrections and clarifications in the new FIGO intrapartum fetal monitoring guidelines. *International Journal of Women's Health, 11*, 1-6. (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6510058/)

2. Bailey, R. E. (2009). Intrapartum fetal monitoring. *American Family Physician, 80*(12), 1388-1396. (https://www.aafp.org)

3. Cervera, A., & Hanzel, L. (2020). Machine learning in healthcare: Benefits, challenges, and future prospects. *Journal of Advances in Applied Healthcare Management, 2*(1), 38-50. https://research.tensorgate.org/index.php/JAAHM/article/view/38

4. Liu, J., & Zhang, X. (2023). Advances in cognitive radio networks: A comprehensive survey. *Journal of Computer Science Research, 5*(2), 1-15. https://journals.bilpubgroup.com/index.php/jcsr/article/view/6242

5.Andrew Mvd. (2020). Fetal health classification dataset. Kaggle. https://www.kaggle.com/datasets/andrewmvd/fetal-health-classification/data

6. Maulik, D., & Mundy, D. (2000). Fetal growth restriction: Pathogenesis and clinical management. *Obstetrics and Gynecology Clinics of North America, 27*(4), 799-825. https://pubmed.ncbi.nlm.nih.gov/16721103/

7. Ayres-de-Campos, D., Bernardes, J., Garrido, A., Marques-de-Sá, J., & Pereira-Leite, L. (2000). SisPorto 2.0: A program for automated analysis of cardiotocograms. *Journal of Maternal-Fetal Medicine, 9*(5), 311-318. https://pubmed.ncbi.nlm.nih.gov/11132590/