Fetal Health Classification

1st Anish Jain
Roll number: 2022077

2nd Dhawal Bansal
Roll number: 2022159

I. Problem Statement

Our project aims to apply statistical machine-learning techniques to classify fetal health based on features extracted from Cardiotocogram (CTG) exams. Fetal health classification is a critical task in obstetrics, aiding in early detection and intervention of potential issues during pregnancy.

The problem we're addressing involves multi-class classification, with the dataset categorized into three classes: Normal, Suspect, and Pathological, as labeled by domain experts. Our objective is to develop a robust model capable of accurately predicting the fetal health class based on the features extracted from CTG exams.

II. Literature Review

Fetal health classification has gained prominence due to its significant implications for prenatal care. The advent of machine learning has brought about a transformative shift in this field, from traditional manual diagnostics to automated systems based on Cardiotocography (CTG) data. Multi-class classification, a challenging aspect of machine learning, plays a pivotal role in distinguishing among categories like 'Normal', 'Suspect', and 'Pathological' fetal states.

Historically, simpler statistical models such as logistic regression were employed, which, while effective for binary outcomes, are less so for multi-class problems without extensive modification. The introduction of k-Nearest Neighbors (k-NN) and decision trees provided more natural extensions to multi-class scenarios but often lacked the precision necessary for medical diagnostics.

Recent studies have emphasized the importance of advanced ensemble methods, such as Random Forests and Gradient Boosting Machines (GBM), which improve prediction accuracy and robustness by aggregating results from multiple models. For example, Breiman's Random Forest algorithm has demonstrated high accuracy in medical datasets by mitigating variance and bias, crucial for applications where misclassifications can have serious consequences.

Moreover, the challenge of imbalanced datasets in fetal health, where 'Normal' cases vastly outnumber 'Pathological' ones, has directed recent research towards synthetic minority over-sampling techniques (SMOTE) and adaptive boosting (AdaBoost) to enhance the sensitivity towards minority classes. Additionally, deep learning approaches, particularly Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), have been explored for their ability to capture temporal patterns and dependencies in sequential CTG data, offering promising results despite their complexity and computational demands.

Despite these advancements, issues such as data privacy, interpretability of machine learning models, and the integration of automated systems into clinical workflows remain challenging. The balance between model complexity and interpretability, especially in a clinical context, is crucial for gaining trust and actionable insights from end-users such as obstetricians and nurses.

In conclusion, the landscape of fetal health classification is evolving, with a clear trend towards more sophisticated, data-driven approaches that promise to enhance the accuracy and timeliness of prenatal diagnostics. The ongoing research continues to push the boundaries of what machine learning can achieve in this vital area of healthcare.

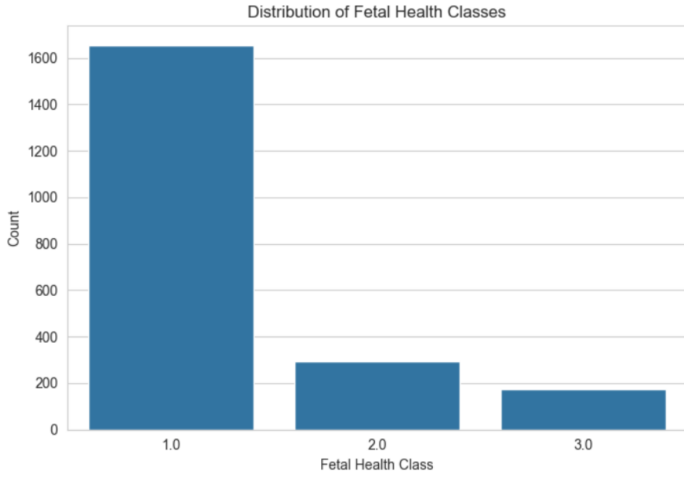
III. Dataset Details

The dataset utilized in this study consists of 2126 records derived from Cardiotocography (CTG) exams, which are used to assess fetal health by monitoring fetal heart rate and other physiological parameters during pregnancy. The CTG data was meticulously recorded and labeled by medical experts into three classes: Normal, Suspect, and Pathological.

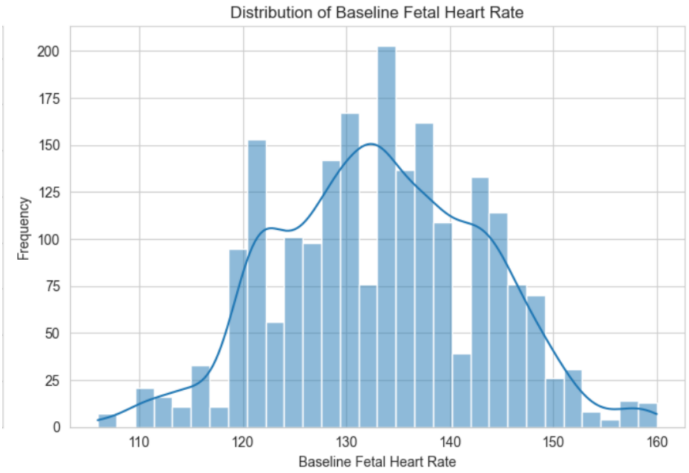
Each record in the dataset contains features such as fetal heart rate, uterine contraction patterns, and various other biophysical measures. These features are crucial in determining the well-being of the fetus and are used by clinicians to make informed decisions regarding the course of pregnancy management.

To validate the robustness of our models and simulate real-world applicability, the dataset was evenly split into two distinct parts:
- Dataset Part 1: Used to train and test the initial set of models. This helped in determining the baseline performance and understanding the behavior of different algorithms under similar conditions.
- Dataset Part 2: Used as a completely independent test set to validate the findings from Dataset Part 1. This approach mimics a scenario where models are deployed in a clinical

setting, facing data from different patient groups.



Distribution of Fetal Health Classes



Distribution of Baseline Fetal Health Rate

IV. Proposed Architecture

A. Initial Approach and Iterations

The modeling began with simpler statistical techniques such as Logistic Regression, serving as a baseline for performance metrics. As the complexity of data understanding increased, more robust models like Random Forest and Support Vector Machines (SVM) were introduced, followed by explorations into Neural Networks. This progression allowed for a comparative analysis of how each model managed the inherent complexities of the dataset, such as non-linear relationships and class imbalances

B. Decision Factors

Several key factors influenced the selection of models throughout the project:

1) Feature Correlations: Preliminary analysis using heatmaps of the dataset indicated varying degrees of correlation between features. Models like Random Forest and Neural Networks, which can handle multicollinearity effectively, were favored for their ability to capture complex interactions between features.

2) CLass Imabalance: The dataset displayed a significant imbalance in the distribution of target classes, which affects model performance, particularly for minority classes. Techniques and models that could incorporate handling of imbalanced data, such as weighted classes in SVMs and ensemble methods in Random Forests, were prioritized.

C. Challenges and Modifications

One significant challenge was the initial attempt to use Principal Component Analysis (PCA) to reduce the dimensionality of the dataset. While PCA helped in reducing the feature space, it did not significantly enhance the model's predictive accuracy when compared to using original features. This could be attributed to the loss of critical information that was not captured by the principal components. Additionally, models tested on PCA-reduced data did not perform as well in distinguishing between the more nuanced 'Suspect' and 'Pathological' classes, leading to the decision to revert to using the full feature set in subsequent models.

D. Further challenges included

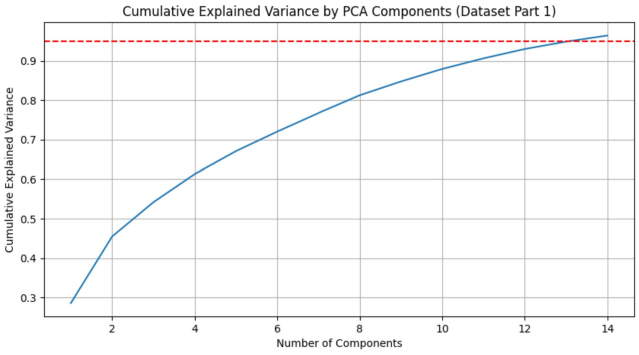
1) Model Convergence: Some of the more complex models, particularly deep neural networks, faced issues with convergence, necessitating adjustments in their architecture and training parameters.
s

2) Computational Efficiency: Given the computational demand, especially with models like SVMs and deep neural networks, there was a continuous need to balance between model accuracy and practical deployment capabilities.

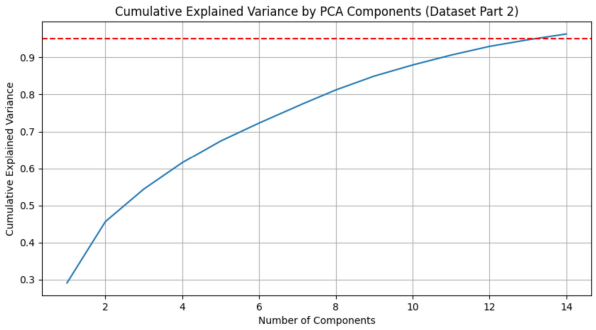
E. Final Model Selection

The final model selection was influenced by a combination of performance metrics and computational efficiency. **Random Forest emerged as the leading model due to its superior handling of class imbalance, robustness against overfitting, and its ability to maintain high accuracy across different performance metrics (precision, recall, and F1-score)**. Additionally, Random Forest proved to be computationally efficient relative to its predictive performance, making it suitable for scenarios where quick decision-making is crucial. This decision was solidified after rigorous cross-validation and performance assessment against other models, affirming its capability to generalize well on unseen data.

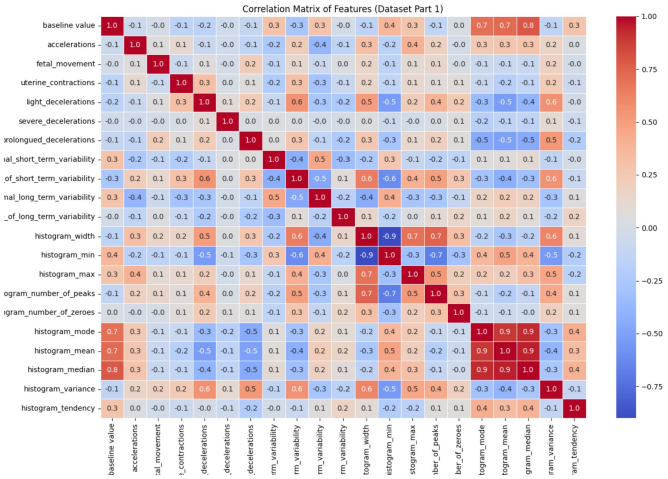
V. VISUALIZATIONS



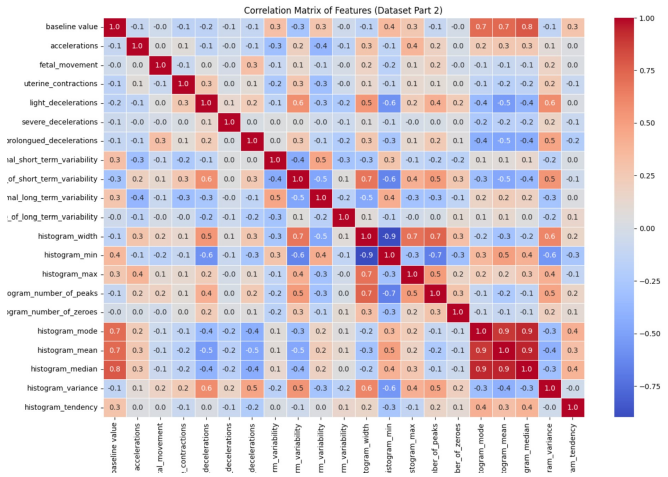
Cumulative Explained Variance (Dataset 1)



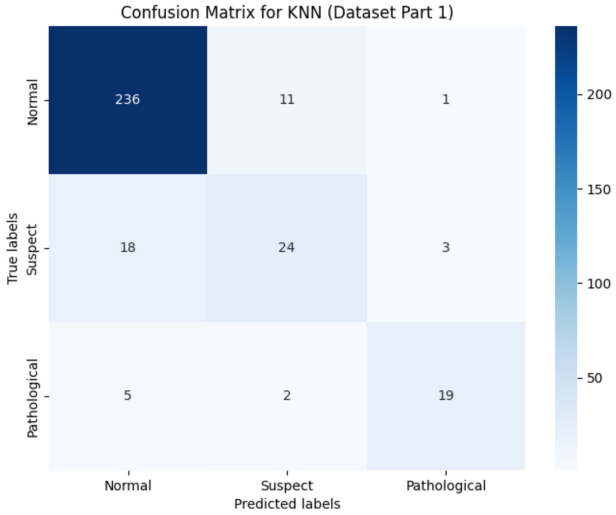
cumulative explained variance (dataset 2)



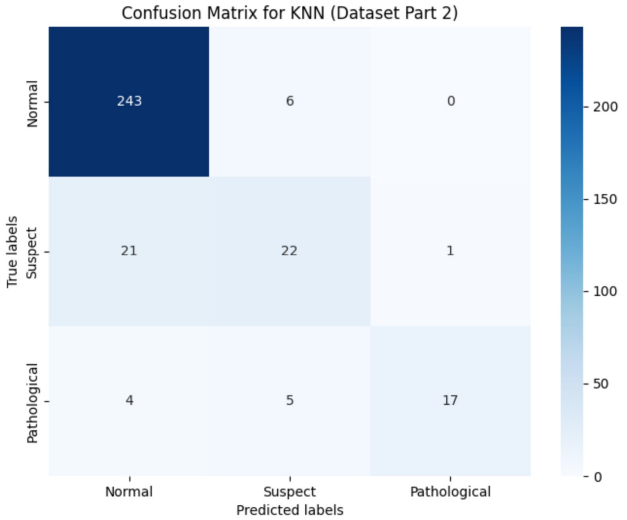
correlation matrix (dataset 1)



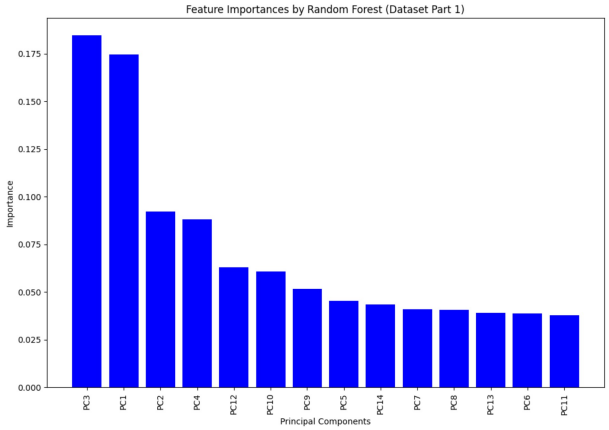
correlation matrix (dataset 2)



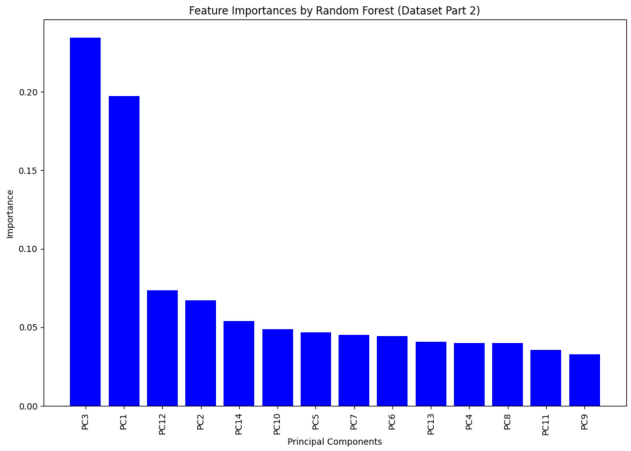
Confusion matrix for KNN (dataset 1)



Confusion matrix for KNN (dataset 2)



Features Importance for Random Forest (dataset 1)



Features Importance for Random Forest (dataset 2)

CLASSIFICATION RESULTS FOR DATASET PART 1

Logistic Regression

Accuracy: 0.8609

Class	Precision	Recall	F1-Score
1.0	0.94	0.93	0.93
2.0	0.65	0.62	0.64
3.0	0.76	0.85	0.80
Macro Avg	**0.78**	**0.80**	**0.79**
Weighted Avg	**0.86**	**0.86**	**0.86**

TABLE I
CLASSIFICATION REPORT FOR LOGISTIC REGRESSION

K Nearest Neighbours

Accuracy: 0.8746

Class	Precision	Recall	F1-Score
1.0	0.91	0.95	0.93
2.0	0.65	0.53	0.59
3.0	0.83	0.73	0.78
Macro Avg	**0.80**	**0.74**	**0.76**
Weighted Avg	**0.87**	**0.87**	**0.87**

TABLE II
CLASSIFICATION REPORT FOR K NEAREST NEIGHBOURS

Random Forest

Accuracy: 0.8966

Class	Precision	Recall	F1-Score
1.0	0.91	0.98	0.94
2.0	0.82	0.51	0.63
3.0	0.88	0.81	0.84
Macro Avg	**0.87**	**0.76**	**0.80**
Weighted Avg	**0.89**	**0.90**	**0.89**

TABLE III
CLASSIFICATION REPORT FOR RANDOM FOREST

Gradient Boosting

Accuracy: 0.9028

Class	Precision	Recall	F1-Score
1.0	0.92	0.97	0.94
2.0	0.78	0.56	0.65
3.0	0.88	0.88	0.88
Macro Avg	**0.86**	**0.80**	**0.83**
Weighted Avg	**0.90**	**0.90**	**0.90**

TABLE IV
CLASSIFICATION REPORT FOR GRADIENT BOOSTING

Support Vector Machine

Accuracy: 0.9028

Class	Precision	Recall	F1-Score
1.0	0.94	0.95	0.95
2.0	0.69	0.69	0.69
3.0	0.88	0.85	0.86
Macro Avg	**0.84**	**0.83**	**0.83**
Weighted Avg	**0.90**	**0.90**	**0.90**

TABLE V
CLASSIFICATION REPORT FOR SUPPORT VECTOR MACHINE

Neural Network

Accuracy: 0.9122

Class	Precision	Recall	F1-Score
1.0	0.94	0.96	0.95
2.0	0.74	0.62	0.67
3.0	0.89	0.92	0.91
Macro Avg	**0.86**	**0.84**	**0.84**
Weighted Avg	**0.91**	**0.91**	**0.91**

TABLE VI
CLASSIFICATION REPORT FOR NEURAL NETWORK

Classification Results for Dataset Part 2

Logistic Regression

Accuracy: 0.8577

Class	Precision	Recall	F1-Score
1.0	0.92	0.96	0.94
2.0	0.63	0.55	0.59
3.0	0.78	0.69	0.73
Macro Avg	**0.78**	**0.73**	**0.75**
Weighted Avg	**0.85**	**0.86**	**0.85**

TABLE VII
CLASSIFICATION REPORT FOR LOGISTIC REGRESSION (DATASET PART 2)

K Nearest Neighbours

Accuracy: 0.8840

Class	Precision	Recall	F1-Score
1.0	0.91	0.98	0.94
2.0	0.67	0.50	0.57
3.0	0.94	0.65	0.77
Macro Avg	**0.84**	**0.71**	**0.76**
Weighted Avg	**0.88**	**0.88**	**0.88**

TABLE VIII
CLASSIFICATION REPORT FOR K NEAREST NEIGHBOURS (DATASET PART 2)

Random Forest

Accuracy: 0.9122

Class	Precision	Recall	F1-Score
1.0	0.92	1.00	0.95
2.0	0.85	0.52	0.65
3.0	0.95	0.77	0.85
Macro Avg	**0.91**	**0.76**	**0.82**
Weighted Avg	**0.91**	**0.91**	**0.90**

TABLE IX
CLASSIFICATION REPORT FOR RANDOM FOREST (DATASET PART 2)

Gradient Boosting

Accuracy: 0.9028

Class	Precision	Recall	F1-Score
1.0	0.91	0.98	0.95
2.0	0.85	0.50	0.63
3.0	0.85	0.85	0.85
Macro Avg	**0.87**	**0.78**	**0.81**
Weighted Avg	**0.90**	**0.90**	**0.89**

TABLE X
CLASSIFICATION REPORT FOR GRADIENT BOOSTING (DATASET PART 2)

Support Vector Machine

Accuracy: 0.8871

Class	Precision	Recall	F1-Score
1.0	0.91	0.98	0.94
2.0	0.69	0.50	0.58
3.0	0.94	0.62	0.74
Macro Avg	**0.85**	**0.70**	**0.76**
Weighted Avg	**0.88**	**0.89**	**0.88**

TABLE XI
CLASSIFICATION REPORT FOR SUPPORT VECTOR MACHINE (DATASET PART 2)

Neural Network

Accuracy: 0.9122

Class	Precision	Recall	F1-Score
1.0	0.94	0.97	0.95
2.0	0.76	0.66	0.71
3.0	0.88	0.81	0.84
Macro Avg	**0.86**	**0.81**	**0.83**
Weighted Avg	**0.91**	**0.91**	**0.91**

TABLE XII
CLASSIFICATION REPORT FOR NEURAL NETWORK (DATASET PART 2)

VI. Inferences and Conclusions

The project focused on applying various machine learning models to classify fetal health based on features from Cardiotocogram (CTG) exams. The objective was to determine the best model based on performance metrics such as precision, recall, and F1-score across three classes: Normal, Suspect, and Pathological.

A. Inferences from Model Performances

Logistic Regression provided a strong baseline with high precision and recall for Class 1 (Normal) but lower scores for Class 2 (Suspect), indicating difficulty in distinguishing less distinct class features. **K Nearest Neighbors** showed improved recall in Class 1 but struggled with precision and recall in Class 2. This suggests that KNN may be sensitive to the noise within the minority classes or the class imbalance. **Random Forest** emerged as the most robust model, achieving the highest marks across almost all metrics, particularly excelling in handling Class 1 and Class 3 (Pathological). Its success can be attributed to its ability to manage class imbalance and its inherent method of averaging multiple decision trees to reduce overfitting and variance. **Gradient Boosting** displayed competitive results, particularly in precision across all classes. However, it slightly lagged behind Random Forest in recall and F1-score for Class 2 and Class 3, which may indicate issues with class imbalance and the model's sensitivity to specific class distributions. **Support Vector Machines** performed consistently across all classes but did not excel in any specific area. This consistency suggests SVMs are reliable but might require further tuning or kernel tricks to handle multi-class nuances better. **Neural Network** showed potential with high precision and recall, particularly in Class 1 and Class 3, reflecting its capability

to capture complex non-linear patterns in the data. However, like other models, it faced challenges with the Suspect class, which could be due to insufficient training epochs or the need for more complex architectures.

The Random Forest model was selected as the final model due to its superior performance and computational efficiency. It proved effective in managing the complexities of the dataset, such as high-dimensional data and class imbalance. The ability of Random Forest to maintain high accuracy and robustness across diverse metrics makes it suitable for practical deployment in clinical settings where quick and accurate decision-making is crucial.

B. Future Directions

Future research should focus on:

- Enhancing the sensitivity of models like SVM and Neural Networks, particularly for underrepresented classes.
- Exploring advanced ensemble techniques and hybrid models that could offer better performance with regard to both accuracy and computational efficiency.
- Integrating machine learning models into real-world clinical workflows to evaluate their effectiveness and user acceptance in practical healthcare environments.

C. Conclusion

The comprehensive evaluation of multiple models underscores the potential of machine learning in enhancing fetal health monitoring. With Random Forest standing out for its robust performance, future advancements in machine learning could further revolutionize prenatal care practices, offering new avenues for improving healthcare outcomes.