Sameedha More, 2023, 11:4
ISSN (Online): 2348-4098
ISSN (Print): 2395-4752

Unveiling the Black Box: Exploring Explainable AI in Predictive Modeling for Fetal Health

Sameedha More, Aarya Patil, Shruti Teltumbade, Saniya Bhalla, Dr. M A Pradhan
Department of Computer Engineering,
AISSMS College of Engineering,
Kennedy Road, Pune- 411001, India

Abstract- This research project aims to develop a predictive model for fetal health using machine learning techniques, with a particular focus on leveraging explainable AI (XAI) methods. The dataset utilized in this study was obtained from Cardiotocography (CTG) machines, providing valuable numerical data related to fetal health monitoring. The project involved pre-processing steps, including oversampling, normalization, and principal component analysis (PCA), to enhance the dataset's quality and relevance. The Random Forest algorithm was selected as the machine learning model due to its classification capabilities and potential for interpretability. Performance evaluation metrics were employed to assess the model's accuracy in predicting fetal health outcomes. XAI techniques were then applied to gain insights into the decision-making process of the model, providing explanations and justifications for its predictions. The outcomes of this study contribute to the prevention of fetal and maternal mortality by providing an interpretable predictive model for early identification of high-risk pregnancies.

Keywords- Machine learning, Random Forest, Explainable AI, Foetal Health.

I. INTRODUCTION

Through this paper, we wish to classify foetal health to prevent maternal and foetal mortality. In 2020, theinfant mortality rate in India was at about 27 deaths per 1,000 births which is a significant decrease from the years prior but still lacking compared to developed countries. Fetal mortality can cause complications in a mother's health. The mother can experience blood clots, heavy bleeding, infection, fever, vomiting and pain if the dead fetus remains in the body resulting in maternal mortality.Most maternal deaths are preventable, as the health-care solutions to prevent or managecomplications are well known.

The accurate classification of fetal health plays a critical role in preventing fetal and maternal mortality. Timely identification of potential risks and complications allows healthcare practitioners to intervene and provide appropriate care, ultimately improving outcomes for both the mother and the unborn child.

In recent years, machine learning (ML) algorithms have shown great promise in assisting medical professionals with early detection and diagnosis. In this paper, we focus on the classification of fetal health using the Random Forest algorithm, selected through an extensive evaluation process using WEKA.

Furthermore, we employ explainable AI techniques to enhance the interpretability of the model's predictions, providing healthcare practitioners with valuable insights into the factors influencing fetal health**.** The outcomes of this study have the potential to assist healthcare professionals in making informed decisions, facilitating early intervention, and improving the overall care provided to expectant mothers and their unborn children.

II. LITERATURE REVIEW

Abolfazl Mehbodniya, Arokia Jesu Prabhu Lazar, Julian Webber et al., [1] "Fetal health classification from cardiotocographic data using

machine learning" In this paper, CTG data is used to classify the health of the fetus using algorithms like multi-layer perceptron, random forest (RF), support vector machine, and K-nearest neighbours. The comparison of results of the algorithms shows that random forest performs best with an accuracy rate of 0.945.

Tomas Peterek, Petr Gajdos, Pavel Dohnalek, and Jana Krohov, [2] "Human Fetus Health Classification on Cardiotocographic Data Using Random Forests" In this work, the authors use Random Forest which performs very well with .9469 accuracy rate.

Nabillah Rahmayanti, Humaira Pradani, Muhammad Pahlawan, Retno Vinarti et al.,[3] "Comparison of machine learning algorithms to classify fetal health using cardiotocogram data" This work compares the following 7 algorithms to predict the fetal health: XG Boost (XGB), Random Forest (RF), Artificial Neural Network (ANN), Light GBM (LGBM), Long-short Term Memory (LSTM), Support Vector Machine (SVM), and K-Nearest Neighbour (KNN). The results show that 5 out of 7 algorithms perform very well (0.89-0.99 accuracy rate). Those five algorithms are RF, SVM, XGB, LGBM, KNN.

Naveen Reddy Navuluri, [4] "Fetal Health Prediction using Classification Techniques" This work proves the prediction accuracy using the different classification models and compares which model performs better. The classification models are random forest, support vector machine (SVM), naïve bayes classifier, and logistic regression. Out of this logistic regression performed better compared to its peers with an accuracy rate of 0.995.

Ramla, M. & S., Sangeetha & Savarimuthu, Nickolas, [5] "Fetal Health State Monitoring Using Decision Tree Classifier from Cardiotocography Measurements" In this paper, the authors use decision tree classifier for implementation and highest accuracy rate is given by gini index which is 0.90.

Pawar, Urja & O'Shea, Donna & Rea, Susan & O'Reilly, Ruairi. [6] (2020). Explainable AI in Healthcare. In this work, Explainable AI is considered as a way to analyze and diagnose health data by AI-based systems and a proposed approach is presented with the aim unveiling the black box of AI.

III. METHODOLOGY

1. Data Preparation:

The dataset employed in this project was obtained from UCI Repository and consisted of numerical data extracted from Cardiotocography (CTG) machines. These machines are widely utilized for monitoring fetal heart rate and uterine contractions during pregnancy.

This dataset contains 2126 records of 22 features extracted from Cardiotocogram reports, which were then classified into 3classes:

- Normal
- Suspect
- Pathological

Table 1. the list of all of the 22 features along with their respective feature number used in the data model.

Feature No.	Feature Name
0	Baseline value (beats per minute)
1	Accelerations (number of accelerationsper second)
2	Fetal_movement (movements of fetal per second)
3	Uterine_contractions (per second)
4	Light_decelerations (per second)
5	Severe_decelerations (per second)
6	Prolongued_decelerations (per second)
7	Abnormal_short_term_variability
8	Mean_value_of_short_term_variability
9	Percentage_of_time_with_abnormal_long_term_variability
10	Mean_value_of_long_term_variability
11	Histogram_width
12	Histogram_min
13	Histogram_max
14	Histogram_number_of_peaks
15	Histogram_number_of_zeroes
16	Histogram_mode
17	Histogram_mean
18	Histogram_median
19	Histogram_variance
20	Histogram_tendency
21	Fetal_health

2. Data Preprocessing:
That is more suitable and effective for machine learning algorithms, data preprocessing steps performed are:

2.1 Missing / Null Values: Null values or missing values negatively affects the accuracy of the result, i.e., they lead to incomplete or biased results. They can be signs of poor-quality data and collection methods, as well as indicate errors or inconsistencies with in the data set. Since we use the fetal health data set, checking for missing values is an essential step to ensure quality, patient safety and making the right therapeutic decisions. We noted that there are no missing or null values within the dataset.

2.2 Over Sampling: In machine learning oversampling is used to address class imbalance in datasets, where one class has significantly fewer sample than the others. In the fetal health classification, the main objective is to predict the health status of the fetus based on various features or attributes.

Oversampling is a technique that help to balance the distribution and improve the model performance by calculating the difference between the number of samples in the majority class and minority class and works by randomly selecting samples from the minority class and creating duplicates until the number of samples in that class matches the number of samples in the majority class.

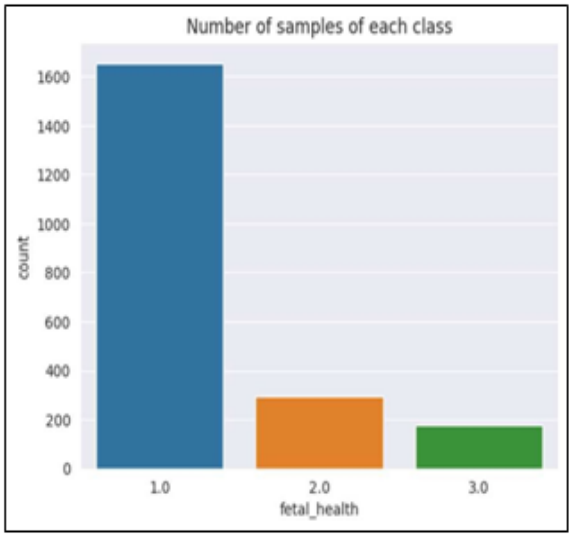

Fig 1. Class count before oversampling.

Here, to balance out the minority classes 2 and 3 that are pathological and suspect respectively with the majority class that is normal in our dataset, random oversampling is used which involves randomly repeating cases of the minority class.

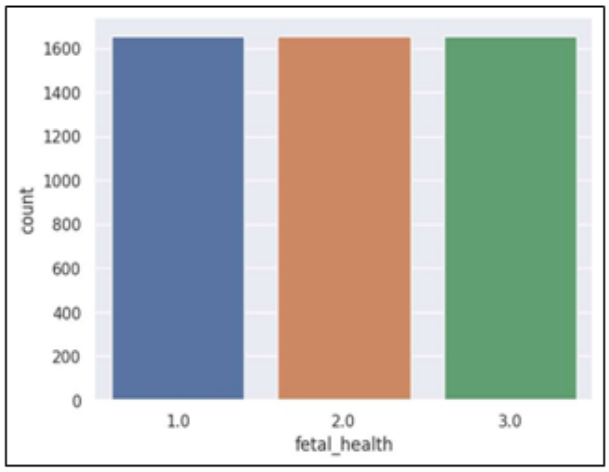

Fig 2. Class count after oversampling.

2.3 Standardization:

- Standardization helps to adjust all the data features to have a common scale to improve the performance. This is done to ensure that no particular feature dominates the learning process.
- The values of dataset features have large scale difference between them.
- For e.g., feature Histogram_max has values greater than 100, while value severe_decelerations are in range 0-1.
- The difference between these two values is large. So,we use scale () function to standardize data.

The Standard Scaler class scales each feature independently by subtracting the mean of the feature and dividing by the standard deviation. The resulting scaled feature has a mean of 0 and a standard deviation of 1.

The mathematical formula for scaling a feature x is:

$$z = (x - mean) / std\_dev$$

2.4 Principal Component Analysis (PCA):
Principal Component Analysis (PCA) streamlines the complexity of high dimensions data while maintaining trends and patterns. To do this, it transforms the data into fewer dimensions, which act as feature summarizer. After using principal

component analysis, it was discovered that all of the models provide greater accuracies than those obtained before to lowering the dimension of the dataset.

To reduce the dimensionality of the data, the PCA algorithm selects the top k eigenvectors with the highest eigenvalues, where k is the desired number of dimensions in the lower-dimensional space. The algorithm then projects the data onto the subspace spanned by these k eigenvectors to obtain the lower-dimensional representation of the data.

The mathematical formula for computing the principal components is:

PC = X * V

Where PC is the matrix of principal components, X is the scaled input data, and V is the matrix of eigenvectors of the covariance matrix of X. The eigenvectors with the highest eigenvalues correspond to the principal components.

After using PCA to the dataset, the number of dimensions is decreased from 22 to 14 principal components while keeping 95% of their variance. As well as we note that there has been an increase in the accuracy, precision, recall, and f1 scores of each of the models.

IV. MODEL ARCHITECTURE

1. Random Forest:
It is a classification algorithm which consists of many decision trees. It is called so because we select random subsets of data and features and therefore end up building a forest of decision trees. It adds more randomness to our model and hence gives better results than decision trees. It uses the training data so that it can learn to make predictions.

2. Gradient Boosting:
It is used in classification and regression tasks. Gradient boosting provides us with predictive accuracy and lots of flexibility. It provides us with many hyper parameter tuning techniques. We use gradient boosting when we want to reduce the bias error.

3. Support Vector Machine (SVM):
SVM is the set of supervised learning based on statistical learning theory use for classification and regression challenges because it has high accuracy and good promotion The main aim of the SVM is to create the optimal hyperplane also referred as decision boundary that can be separate by 2 or moreclasses so that new data points can be put in correct class.

4. Support Vector Machine (RBF):
The RBF kernel is not a parametric model. It is used as a default kernel within the sklearn's SVM classification algorithm. The RBF kernel shows similarity to the K-Nearest Neighbour Algorithm. It isa nonlinear classifier.

5. Decision tree:
It is a popular and powerful supervised learning algorithm which is utilized for classification as well as regression tasks. It is a decision support tool that usethe concept of trees to structure the given information in the sequence of decision and consequence.

6. K-nearest neighbor:
K-nearest neighbor is a non-parametric, sophisticated, and one of the simple supervised machine learning techniques which is used to solve classification using k numbers of neighbors.

7. Logistics Regression:
It is the type of classification algorithm that comes under supervised machine learning. Logistics regression is the statistical algorithm used for binary classification problems. It uses the logistic function (also called the sigmoid function) to map the linear combination of the independent variables to a probability between 0 and 1.

After observing the performance of each machine learning algorithm, it was observed that random forest gave the best results.

V. MODEL EVALUATION

1. Confusion Matrix
A confusion matrix is a useful tool for evaluating the performance of a multi-class classification model. It is a square matrix that compares the predicted labels ofthe model with the true labels of the data.

Let's assume we have a multi-class classification problem with N classes (e.g., Class 1, Class 2, ..., Class N). The confusion matrix for this problem would be an N x N matrix. Here's an example of a confusion matrix for a multi- class classification problem with three classes (Class 1, Class 2, and Class 3):

In the confusion matrix:

- **TP (True Positive):** The number of instances correctly predicted as belonging to a particular class.
- **FP (False Positive):** The number of instances incorrectly predicted as belonging to a particular class when they actually belong to a different class.
- **FN (False Negative):** The number of instances incorrectly predicted as not belonging to a particular class when they actually belong to that class.
- **TN (True Negative):** The number of instances correctly predicted as not belonging to a particular class.

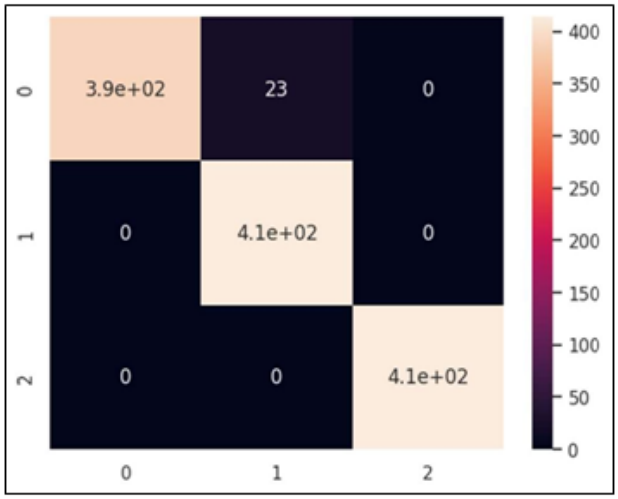

Fig 3. Confusion Matrix.

Summarizing the confusion matrix for each class:

- Class 1: TP1 instances are correctly predicted as Class 1, FP1 instances are incorrectly predicted as Class 1, and FN1 instances are incorrectly predicted as not Class 1.
- Class 2: TP2 instances are correctly predicted as Class 2, FP2 instances are incorrectly predicted as Class 2, and FN2 instances are incorrectly predicted as not Class 2.
- Class 3: TP3 instances are correctly predicted as Class 3, FP3 instances are incorrectly predicted as Class 3, and FN3 instances are incorrectly predicted as not Class 3.

The confusion matrix provides valuable insights into the model's performance, such as accuracy, precision, recall, and F1-score for each class. These metrics can help in evaluating the model's ability to classify instances correctly across all classes.

2. Receiver Operating Characteristic Curve (ROC Graph):

In machine learning, the ROC (Receiver Operating Characteristic) curve is a graphical representation which illustrates the performance of a classification model. It displays the trade-off between the true positive rate (TPR) and the false positive rate (FPR) for the various classification thresholds.

The ROC curve provides a visual representation of the model's ability to discriminate between the two classes. A perfect classifier would have a curve that passes through the top-left corner of the plot (TPR = 1, FPR = 0), indicating high true positive rates and low false positive rates for all classification thresholds. The closer the curve is to the top-left corner, the better the model's performance.

Additionally, the area under the ROC curve (AUC- ROC) is often used as a metric to quantify the overall performance of the model. A perfect classifier would have an AUC-ROC of 1, while a random or non- informative classifier would have an AUC-ROC of 0.5.

In the case of a multiclass classification, we use the following method:

3. One VS rest:

It is the method to evaluate the multiclass models by comparing each class with all the others at the same time. In this method, we take one class and consider it as our positive class, while the rest are considered as the negative class. By examining the ROC curve and the corresponding AUC-ROC value, you can evaluate and compare the performance of different classification models and choose the one that best suits your requirements.

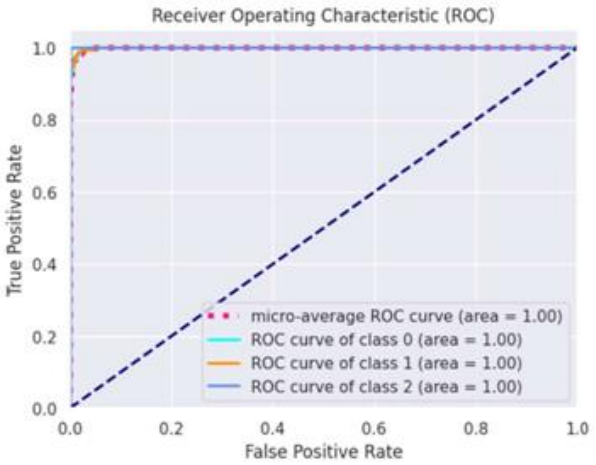

Fig 4. Receiver Operating Characteristic Curve (ROC Graph).

VI. EXPLAINABLE AI

Explainable AI (XAI) is an Enhancement of artificial intelligence systems that provides a simple, clear, and easily understandable explanation of the decision-making processes. XAI is an emerging field within AI research, which aims to address the "black box" problem of traditional machine learning models, where the decision-making process is opaque to human understanding.

In traditional machine learning models, complex algorithms are used to make decisions and those decisions are difficult to explain to humans. This can lead to a lack of transparency, which can be problematic in healthcare scenarios where the consequences of a decision are significant.

XAI seeks to overcome this challenge by developing AI models that can provide human-understandable explanations of their decision-making processes. By providing clear explanations of their decision-making processes, XAI models can help increase transparency and trust in AI systems.

1. SHAPASH:

Shapash is an open-source Python library that helps users to create interpretable and interactive dashboards for their machine learning models.

It was developed by the French consulting firm, MAIF, and is designed to make it easy for data scientists and developers to share the results of their machine learning models with business stakeholders. Shapash uses the SHAP (SHapley Additive explanations) framework to provide model interpretability. It generates summary statistics and visualizations that show the impact of each feature on the model's predictions, helping users to understand how the model works and to identify potential biases or errors.

Shapash can be used for regression as well as classification models. It also provides an interactive dashboard that allows users to explore the model's predictions and to drill down into specific subsets of the data. The dashboard can be customized to include specific visualizations and metrics, making it easy to communicate the results of the model to business stakeholders.

2. Our work on XAI:

When building an ML model to predict fetal health, it is important to ensure that the model's decisions are understandable and transparent to medical professionals and patients. The use of XAI in our project enables us to provide clear and interpretable explanations of how the ML model arrived at its decision.

XAI in our project could be used to provide doctors with more transparent and interpretable medical diagnoses, helping to improve patient outcomes. We have used SHAPASH XAI library to explain the results to the end users. Using XAI SHAPASH we have created a web application/ dashboard which explain the feature contribution for our predictions.

Fig 5. XAI- Shapash Dashboard.

3. Feature Importance:
The feature importance plot displays the importance or impact of each feature on the model's predictions. It ranks the features based on their relevance in influencing the model's output. The plot helps you understand which features have the most significant influence on the model's predictions.

This information can be valuable for feature selection, understanding the model's behavior, and identifying potential areas for improvement. This plot represents the bar plot indicating the contribution of each feature for predicting the fetal health.

In the figure, we can observe that the "histogram_mean:" feature contributed the most for the classification.

4. Feature Contribution:
The feature contribution plot shows the contributionof each feature to the prediction of a specific sample or instance. It breaks down the prediction into individual feature contributions, allowing you to understand how changes in each feature affect the overall prediction. In the figure, the scatter plot shows the prediction probability for histogram_mean feature.

5. Dataset:
The dataset component of the Shapash web dashboard provides an overview of the input data used for training and inference. It typically includes the feature values for each sample or instance in the dataset and allows you to explore the data, examine individual samples, and potentially identify patterns or outliers that may impact the model's predictions.

6. Local Explanation:
The local explanation feature in Shapash provides detailed explanations for individual predictions madeby the model. It enables you to identify which features played a significant role in a specific prediction and how positively and negatively they influenced the model's decision-making process.

In the figure, features contributing positively for a particular prediction are shown on the right side of the line in yellow and those contributing negatively are on the left side in blue.

VII. CONCLUSION

This research contributes to the objective of preventing fetal and maternal mortality. Comprehensive approach to fetal health prediction using a fetal dataset is presented in this paper. Through various pre-processing techniques and the selection of the Random Forest algorithm, we achieved accurate predictions of fetal health outcomes.

The application of explainable AI techniques allowed for insights into the model's decision-making process, aiding in understanding and trust-building. Implementing this model can enhance healthcare outcomes.

However, limitations such as data availability and model interpretability should be addressed. Future research should focus on expanding the dataset, extracting data from CTG graphs, exploring alternative algorithms, and improving interpretability. Overall, this study emphasizes the value of predictive modelling in improving fetal health and healthcare decision- making.

REFERENCES

[1] M Mehbodniya, A., Lazar, A. J. P., Webber, J., Sharma, D. K., Jayagopalan, S., K, K., Singh, P., Rajan, R., Pandya, S., &Sengan, S. (2021). Fetal health classification from cardiotocographic data using machine learning. Expert Systems, e12899. https //doi.org/10.1111/exsy.12899

[2] P. Tomáš, J. Krohová, P. Dohnálek and P. Gajdoš, "Classification of cardiotocography records by random forest," 2013 36th International Conference on Telecommunications and Signal Processing (TSP), Rome, Italy, 2013, pp. 620-923, doi: 10.1109/TSP.2013.6614010

[3] Rahmayanti, Nabillah Annisa et al. "Comparison of machine learning algorithms to classify fetal health using cardiotocogram data." Procedia Computer Science (2022)

[4] Naveen Reddy Navuluri, 2021, Fetal Health Prediction using Classification Techniques, INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH & TECHNOLOGY (IJERT) Volume 10, Issue 11 (November 2021)

International Journal of Science, Engineering and Technology

An Open Access Journal

[5] Ramla M. & S., Sangeetha & Savarimuthu, Nickolas. (2018). Fetal Health State Monitoring Using Decision Tree Classifier from Cardiotocography Measurements. 1799- 1803. 10.1109/ICCONS. 2018.86630

[6] Pawar, Urja & O'Shea, Donna & Rea, Susan & O'Reilly, Ruairi. (2020). Explainable AI in Healthcare. 10.1109/CyberSA49311.2020.913 96 55.