

Article

An Explainable Hybrid TabNet–Residual MLP Framework for Robust Fetal Health Prediction Using Focal Loss and Leakage-Aware Cross-Validation

Samaa Ahmed *, Doaa Saad  and Ahmed Yakoub 

Information Systems Department, Faculty of Computers and Artificial Intelligence, Capital University (Formerly Helwan University), Cairo 11795, Egypt; doaa.saad@fci.capu.edu.eg (D.S.); ahmedYakoub@fci.capu.edu.eg (A.Y.)

* Correspondence: samaaahmed252@fci.capu.edu.eg

Abstract

Effective fetal health prediction is crucial for early diagnosis of fetal distress and avoiding negative perinatal effects. Cardiotocography (CTG), which measures fetal heart rate and uterine contractions, is a popular method for prenatal examination; however, conventional interpretation is arbitrary and unreliable. Fetal health prediction continues to face challenges due to scarce and imbalanced CTG datasets and data leakage during model evaluation, which can lead to poor generalization and inaccurate performance estimates. Moreover, the absence of explainable AI reduces model clarity and restricts clinical utilization. This study offers a hybrid deep learning model that uses TabNet, Residual Multi-Layer Perceptron (Residual MLP), and Focal Loss to classify normal, suspect, and pathological fetal states. TabNet allows for attention feature learning from CTG data, Residual MLP increases predictive robustness, and Focal Loss aids minority abnormal case diagnosis. To achieve a reliable evaluation, data splitting is used to create a pipeline designed to eliminate conventional train–test data leakage by performing data partitioning before model training and evaluation. SHAP and LIME enhance interpretability by offering clear global and local explanations. The suggested model obtains 99.30% accuracy, 99.10% balanced accuracy, and 98.65% pathological recall on the augmented dataset. A comparative evaluation shows that fixing data leakage drops overinflated baseline accuracy from 96.80% to 93.14%, emphasizing the necessity of a robust experimental design. The results show that the proposed framework outperforms cutting-edge fetal health prediction approaches, offering a robust, understandable, and clinically reliable alternative for CTG-based fetal health evaluation.

Keywords: fetal health monitoring; cardiotocography (CTG); TabNet; data leakage-free validation; explainable artificial intelligence (XAI)



Academic Editor: Hongying Meng

Received: 3 July 2026

Revised: 27 July 2026

Accepted: 28 July 2026

Published: 5 August 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Reliable fetal health monitoring is a fundamental component of prenatal and intrapartum treatment, since early discovery of fetal anomalies can significantly reduce neonatal morbidity, death, and long-term neurological disorders [1–3]. Fetal distress, which is frequently associated with insufficient oxygen supply or maternal problems such as hypertension, can result in serious adverse consequences such as preterm birth, cerebral palsy, developmental abnormalities, and even fetal mortality if not detected and treated promptly [4,5]. As a result, continuous fetal monitoring has become a crucial clinical practice during pregnancy and labor [6].

Cardiotocography (CTG) is one of the most commonly utilized non-invasive fetal monitoring techniques since it measures both fetal heart rate (FHR) and uterine contractions. CTG provides valuable physiological data regarding fetal oxygenation and autonomic nervous system function by analyzing temporal changes in fetal heart activity and contraction patterns, allowing clinicians to detect irregularities like tachycardia, cardiac bradycardia, fetal hypoxia, and pathological decelerations [7]. Modern CTG files contain several quantitative parameters, such as baseline fetal beat, accelerations, decelerations, variability measurements, fetal movements, and uterine contractions, which combined provide a structured tabular dataset suited for automated fetal health prediction [8–11]. Despite its clinically significant importance, CTG interpretation remains highly subjective and heavily dependent on clinician expertise, frequently resulting in multi-observer variability, ambiguous diagnoses, and high false-positive rates, which can contribute to unnecessary medical interventions or missed pathological cases [12].

To enhance the trustworthiness of CTG analysis, artificial intelligence (AI), especially machine learning (ML) and deep learning (DL), has been actively researched for automated fetal health classification. Standard ML techniques, such as Decision Trees (DTs), Support Vector Machines (SVMs), Random Forest (RF), and ensemble learning methods, have shown promising results in identifying prenatal health patterns. Deep learning models have recently demonstrated greater capability in predicting complicated nonlinear interactions and creating distinguishing characteristics directly from clinical data, which reduces the need for human feature engineering and enables more reliable decision support systems [13,14].

Meanwhile, some significant issues remain unsolved. Public CTG databases are typically limited and unevenly distributed, with challenging fetal cases severely under-represented. This mismatch frequently biases prediction models towards the majority class, lowering their ability to adapt to clinically significant abnormal instances [15,16]. Furthermore, several recent studies use data augmentation before partitioning the dataset into training and testing sets, resulting in data leakage that can significantly inflate claimed performance while reducing real-world generalizability. Additionally, the poor comprehension of many deep learning models prevents their use in medical settings, where clear and reliable predictions are critical.

To overcome these challenges, this study proposes a new fetal health prediction framework that integrates an LLM-aided rule-based synthetic data generation methodology and a hybrid deep learning architecture. Unlike traditional synthetic data generation methods, which rely primarily on statistical oversampling or deep generative models, the proposed methodology uses a Large Language Model (LLM) controlled by medically established rules and class-conditioned constraints to generate physiologically realistic CTG samples while maintaining feature relationships and class characteristics. This transparent and reproducible generating procedure creates a massive, medically consistent synthetic dataset that can be used to train and evaluate AI models for prenatal health prediction. The prediction approach combines TabNet, which uses sequential attention to identify the most relevant clinical attributes in tabular data, with a Residual Multi-Layer Perceptron (Residual MLP), which refines previously learnt representations. Furthermore, Focal Loss is added to increase the recognition of rare problematic situations, while SHAP and LIME are utilized to propose both global and local explanations, improving model interpretability and aiding reliable medical decision-making.

The fundamental scientific contributions of this research are outlined as follows:

- A novel LLM-aided rule-based synthetic data generation approach is recommended to generate medically appealing CTG samples while maintaining feature dependencies and class distributions.

- A large, medically consistent CTG dataset is created using a transparent and repeatable synthetic data generation methodology. This provides a scalable alternative to standard analytical oversampling and deep generative methodologies for medical tabular data.
- A hybrid TabNet–Residual MLP architecture tuned with Focal Loss is proposed to enhance fetal health prediction, with a special emphasis on improving the classification of rare problematic situations.
- A comprehensive experimental evaluation is performed utilizing stratified cross-validation and several performance indicators to establish the robustness, efficacy, and applicability of the proposed framework.
- Explainable AI techniques (SHAP and LIME) are combined to provide all global and local model interpretations, thus increasing transparency and facilitating reliable clinical decisions.

Overall, the proposed methodology combines medically guided LLM-aided synthetic data generation, an improved hybrid deep learning prediction model, and explainable AI to create a unified strategy for intelligent prenatal health prediction. By addressing key limitations identified in previous studies, such as limited data availability, class imbalance, and methodological issues like data leakage, the proposed framework creates a more robust, clear, and clinically relevant AI-assisted decision support system for obstetric medical care.

The subsequent sections of this paper are arranged as follows. Section 2 provides a complete literature overview of fetal health prediction, with a focus on cardiotocography (CTG)-based machine learning and deep learning algorithms, as well as current problems and methodological constraints. Section 3 explains the suggested research methodology, which includes data preprocessing, synthetic augmentation, leakage-free validation, a hybrid TabNet–Residual MLP architecture, and an evaluation mechanism. Section 4 summarizes the suggested model’s experimental results, including comparative performance evaluation and explainability outcomes based on SHAP and LIME. Section 5 reviews the findings, emphasizing their predictive power and methodological significance, and wraps up the study by reviewing the important contributions, limitations, and prospective future research areas.

2. Related Works

This section examines and evaluates the collection of research on computational intelligence-based fetal health prediction. The development and construction of prediction models, including both conventional machine learning and advanced deep learning techniques, are given special consideration, as are the kinds and attributes of datasets used. The efficacy and dependability of the suggested approaches are also evaluated by inspecting the performance indicators and evaluation techniques described in the studies. The review also seeks to identify common limitations, difficulties, and research gaps while highlighting the advantages and contributions of contemporary methods.

Ahmed and Mahmoud [17] constructed an AI-based model utilizing the UCI Cardiotocography (CTG) dataset (2126 samples, 21 characteristics) to predict fetal health early. It employed feature selection (including mutual information, chi-square, correlation, and PCA), preprocessing (outlier removal, normalization, and standardization), and the Synthetic Minority Oversampling Technique (SMOTE) for handling the class imbalance problem. Grid Search was then used for hyperparameter optimization. Different machine learning and deep learning models were analyzed, including SVM and ANN, with ensemble approaches outperforming; the stacking classifier achieved 98.9% accuracy with high precision, recall, F1-score, and AUC. However, the study is constrained by employing a single dataset, high computing cost, no result explainability, and possible data leakage caused

by preprocessing and augmentation before train–test splitting, which might overestimate actual performance.

Feng et al. [18] developed a sophisticated artificial intelligence-driven framework for fetal health classification utilizing cardiotocography (CTG) data. The paper established a stacked ensemble learning strategy to increase predictive performance by successfully merging numerous base learners and also integrating kernel-based SHAP to enhance model clarity by measuring each attribute's impact on the final prediction. The suggested model was tested using CTG data and attained an accuracy of 95.39%, indicating a beneficial ability to discriminate between normal, suspect, and abnormal fetal situations. Furthermore, due to its outstanding predictive accuracy, the framework's use of explainable AI techniques increased clinical realism, making it more acceptable for medical decision support. Nevertheless, the research was constrained by the dataset's modest size, which could limit the model's capacity to generalize successfully to broader and more diverse medical groups, therefore compromising its resilience in practical applications.

Barnova et al. [19] provided an extensive evaluation of the use of artificial intelligence and machine learning approaches in electronic fetal monitoring devices. It explained the way fetal heart rate impulses and related biological information were processed using several AI-based algorithms to aid in noise elimination, feature extraction, and fetal health classification. This research found that machine learning and deep learning models are increasingly effective at enhancing the accuracy and dependability of fetal status examinations when compared to older approaches. It also evaluated major datasets, techniques, and mathematical trends in this field, highlighting AI's potential to improve medical judgment in obstetrics. Even so, the paper recognized a number of problems, including data quality problems, inadequate generality between groups, and the demand for more accurate and interpretable models for practical applications.

Chiou et al. [20] employed a cardiotocography (CTG) dataset from 552 individuals to construct a deep learning-powered prediction framework for early diagnosis of fetal abnormalities using a convolutional neural network (CNN). The suggested method concentrated on evaluating various real-life labeling approaches, including comparing objective clinical metrics (like umbilical cord blood pH) against subjective evaluations (Apgar grades), while also adding preprocessing and hyperparameter modifications to improve model performance. The findings showed that networks trained on objective pH-sensitive labels made more consistent, accurate, and robust predictions, even during time-based distribution shifts, emphasizing the need for reliable labeling in healthcare AI applications. However, the research is limited by the relatively small dataset size, reliance on a single medical source, potential lack of universality across groups, and sensitivity to temporal variability in CTG signals, which may affect the reliability and realistic applicability of the prediction model.

Cao et al. [21] developed a multimodal deep learning system for fetal health monitoring, combining cardiotocography (CTG) signals and maternal medical records, utilizing a convolutional neural network (CNN) for feature extraction and fusion. The suggested methodology outperforms single-source methods in prediction performance by combining CTG signals with clinical characteristics. The LGBM classifier successfully addressed class imbalance, obtaining an F1-score of 0.9376 for normal cases and 0.8223 for abnormal instances, according to the data, while the model attained an accuracy of 90.77% with an AUC of 0.9201. Nevertheless, the study's limitations include reliance on particular clinical datasets, inconsistent data quality, increased model complexity as a result of multimodal fusion, and possible difficulties with applicability in various clinical settings.

Wong et al. [22] proposed a novel prediction model based on Kolmogorov–Arnold Networks (KANs) for fetal health categorization, using a publicly available cardiotocography

(CTG) dataset containing statistical variables derived from fetal heart rate recordings. The suggested approach offers an alternative to conventional neural networks by leveraging KANs' powerful nonlinear modeling capabilities to capture intricate correlations in CTG data. The outcomes showed that the model outperformed traditional machine learning techniques, achieving classification accuracies of 93.6% for binary classification and 92.6% for three-class classification. Even so, the study's limitations include its dependence on a single standard dataset, the absence of actual medical validation, and possible difficulties with model interpretability brought on by the complicated nature of KAN design, which could limit its use in clinical judgment.

Swain et al. [23] employed a specialized dataset of 1500 fetal health records from a single medical institution in Odisha, India, and proposed a multi-stage pipeline that combined Spectral Clustering, PCA for dimensionality reduction, and two ensemble approaches: a bagging ensemble based on Random Forest (RF), Logistic Regression (LR), and Gradient Boosting, optimized using GridSearchCV, and a voting ensemble of Decision Tree (DT), Support Vector Machine (SVM), and Naïve Bayes (NB). With the highest accuracy of 94.02%, the suggested meta-model that combined DT and RF outperformed similar current models. Nevertheless, the analysis's limited one-center dataset, lack of external validation, and absence of significant medical evaluation parameters like sensitivity, specificity, and AUC-ROC restrict the approach's universality and clinical benefit.

Akbulut et al. [24] employed a custom clinical dataset of 96 pregnant women that was gathered through maternal questionnaires and assessments by three medical professionals in Turkey. This dataset included parameters related to the medical history of the mother and her family to predict the status of fetal anomalies. The suggested architecture combined a binary classification system with an electronic health care mobile application. Nine machine learning techniques were trained and evaluated, and tree-based feature selection was used to determine which clinical characteristics were most distinctive. With an accuracy of 89.5%, an F1-score of 75%, and an AUC of 95%, the Decision Forest model performed the best. Still, the study's practical relevance and generalizability are severely constrained by its very small sample size, primary reliance on the mother's clinical history rather than direct fetal physiological signals like CTG data, and its binary classification framework, which fails to properly represent the multi-class nature of fetal health outcomes.

Vani [25] presented a deep learning-driven medical decision support system that uses cardiotocography (CTG) data to classify the health state of fetuses. By giving input variables varying degrees of relevance, the study created a weighted deep neural network that enhances classification performance and facilitates more efficient learning from uterine contraction patterns and fetal heart rate. The model demonstrated high predictive potential for fetal health evaluation when tested on CTG data to differentiate between normal, suspected, and abnormal fetal states. The findings supported the suggested weighted neural network's potential application in enhancing maternal decision-making by showing that it increased classification accuracy when contrasted with baseline methods. Nevertheless, the study's dependence on a single dataset was a limitation, and it would be difficult to generalize to larger medical facilities with more varied communities.

Chen et al. [26] investigated the issue of class imbalance in fetal health classification using cardiotocography (CTG) data. Therefore, to address the unequal distribution of normal, suspicious, and pathological instances, the study suggested a cost-aware weighted random forest (WRF) model that allocates distinct class weights. Given an average AUC of 0.99 and an F1-score of 97.56%, the model performed well when tested on the UCI CTG dataset, exceeding a number of current techniques. The findings showed that the identification of minority classes in unbalanced prenatal health data was greatly enhanced

by the use of cost-sensitive learning. However, the study's dependence on a single standard dataset could restrict the model's applicability to a wider range of clinical settings.

Fasihi et al. [27] suggested a shallow 1D-CNN model that uses cardiotocography (CTG) signals to classify fetal health. Manual feature engineering was not necessary because the model automatically learned characteristics from raw CTG data. The suggested method performed well in identifying historical trends in fetal heart rate signals, achieving an accuracy of roughly 97.46% when tested on CTG datasets for multi-class classification (normal, suspicious, and pathological). Despite this, the study was constrained by the size and imbalance of the datasets, which might have an impact on how well the model generalizes in actual clinical settings.

Li and Liu [28] suggested a hybrid ensemble learning technique for classifying fetal health based on cardiotocography (CTG) data. For the purpose of building ensemble systems using soft voting and stacking techniques, the study initially used a variety of machine learning models. The top-performing models were then blended into a blender model. The examination was conducted using the CTG dataset, which comprises 2126 samples that were categorized by experts into normal, suspect, and abnormal categories. With an accuracy of 95.9% and high AUC, precision, recall, F1-score, and MCC values, the suggested blender model outperformed conventional individual models. Nevertheless, the study was constrained by inadequate data augmentation methods, which could result in overfitting and diminishing generalization capacity in practical applications.

Sirisha et al. [29] suggested a Dynamic Multi-Layer Perceptron (DMLP) model for using CTG data to categorize fetal health. SMOTE, feature selection, and scaling were among the preprocessing methods used in the study to enhance model performance. The suggested model outperformed various conventional machine learning techniques, achieving an accuracy of 97% using a dataset of 2126 instances. Meanwhile, the study's dependence on a single data source and a rather small dataset may limit its ability to generalize.

Suganthi and Basilica [30] investigated the extent to which deep learning methods could enhance fetal health evaluation using cardiotocography (CTG) data. The authors examined several neural architectures, such as CNN, ANN, and RBFNN, rather than depending just on one modeling approach. They also improved input quality using sophisticated feature selection techniques like Genetic Algorithm and Particle Swarm Optimization. The experimental findings showed that adjusting feature subsets had a significant impact on model performance using a dataset of 2126 instances; the optimal configuration achieved an accuracy of 96.47%. The results highlighted how critical it is to combine deep learning and intelligent feature selection to increase predictive power. However, the reliance on a small dataset raises questions about the model's resilience in more varied clinical settings.

Salini et al. [31] employed a publicly accessible cardiotocography (CTG) dataset with 2126 records that were clinically classified by three skilled obstetricians into normal, suspect, and pathological categories. The suggested framework evaluated several machine learning classifiers, such as Random Forest, Decision Tree, K-Nearest Neighbor, Logistic Regression, Support Vector Classifier, and Gradient Boosting Classifier, after preprocessing the data using standardization and Pearson correlation-based feature selection. Random Forest achieved the highest accuracy of 93%, outperforming the other methods among the models that were evaluated. However, the study's robustness and generalizability in actual clinical settings are constrained by the utilization of a single relatively small dataset, insufficient treatment of class imbalance, limited evaluation criteria, the lack of explainability techniques, and the lack of external validation.

Nagavibha et al. [32] developed a machine learning-powered structure to facilitate fetal health classification with the goal of increasing the accuracy, performance, and speed

of pregnancy diagnostics. The study used a dataset that included medical information and symptoms of pregnant women, particularly those related to fetal risk factors. Ten machine learning models were assessed, including ensemble-based methods. With an accuracy of 96.9%, LGBM outperformed all other models examined, proving its value for classifying fetal health. Nevertheless, the study mostly uses a single dataset, which would restrict how broadly the findings can be applied to other clinical contexts.

Mehbodniya et al. [33] investigated the classification of fetal states into normal, suspect, and pathological classes utilizing cardiotocography (CTG) data for fetal health prediction. In addition to correlation and regression studies to assess feature relevance, a number of machine learning models, such as Support Vector Machine, Random Forest, Multi-Layer Perceptron, and K-Nearest Neighbors, were used and contrasted. With an accuracy of 94.5%, Random Forest outperformed the other classifiers in all important parameters among the assessed models. The results underlined the significance of feature interactions in enhancing prediction performance and showed how successful CTG-based machine learning techniques are for evaluating fetal health. Additionally, the study proposed that advanced feature engineering may be used to further increase the accuracy and adaptability of the model.

In conclusion, the reviewed studies demonstrate that most current methods for predicting fetal health mainly rely on traditional machine learning techniques, while relatively few studies have investigated deep learning approaches, despite their strong potential to improve predictive performance by capturing complicated temporal and nonlinear patterns in cardiotocography (CTG) signals. Additionally, a number of similar constraints were found throughout the research. Reliance on comparatively limited and frequently single-center datasets is a major problem that limits model robustness and generalizability, especially in an important medical field that demands massive and varied samples. Furthermore, data leakage was found in several studies. This usually happens during incorrect data splitting, which results in overly optimistic performance estimates and lowers the accuracy of the reported results in real-life scenarios. Examples of this are performing preprocessing, feature selection, or oversampling before separating training and testing sets. The low adoption rate of explainable artificial intelligence (XAI) methods is another significant barrier. Because healthcare professionals need clear and understandable predictions to support decision-making, validate model behavior, and guarantee patient safety, the lack of interpretability limits clinical trust. Developing trustworthy, broadly applicable, and clinically beneficial prenatal health prediction systems requires addressing these issues. Table 1 summarizes the current research overview of existing fetal health prediction studies, highlighting the datasets utilized, methodology applied, evaluation metrics employed, obtained performance, and notable limitations.

TabNet, one of the deep learning architectures for tabular data, is especially suitable for CTG-based fetal health prediction because it uses a sequential attention mechanism to dynamically choose the most valuable clinical features. Unlike traditional MLPs, which process all features identically, and CNNs or RNNs, which are optimized for visual and sequential information, TabNet is specifically tailored for organized tabular datasets, delivering great predictive performance and natural interpretability. Compared to popular tree-based models like Random Forest and XGBoost, TabNet offers comprehensive, complete representation training with built-in interpretability. To improve its discriminative capacity, this study integrates TabNet with a Residual Multi-Layer Perceptron (Residual MLP), which learns complementary nonlinear characteristic representations through residual connections. This hybrid design uses the benefits of both models to provide a more accurate, durable, and understandable fetal health prediction.

Table 1. Summary of the Related Works.

Study	Dataset	Techniques	Evaluation Metrics	Highest Performance	Explainability Techniques	Limitations
Ahmed and Mahmoud [17]	Cardiotocography (CTG) dataset	KNN, SVM, RF, ET, LGBM, Voting, Stacking, RNN, MLP, LSTM, ANN, TabNet	Accuracy, Precision, Recall, F1-score, AUC	Accuracy = 98.9%	No	Small dataset size, data leakage, and high computing cost
Feng et al. [18]	Cardiotocography (CTG) dataset	SWM, RF, XGB, Stacking	Accuracy, Precision, Recall, F1-score	Accuracy = 95.39%	SHAP	Limited Data Size
Barnova et al. [19]	CTG, fetal electrocardiography (fECG), fetal phonocardiography (fPCG), fetal magnetocardiography (fMCG) databases	RF, SVM, DT, ANN, CNN	Accuracy, Specificity, Sensitivity, F1-score, ROC-AUC	Not Stated	No	Signal Complexity, Lack of Standardization
Chiou et al. [20]	Cardiotocography (CTG) dataset	CNN	ROC-AUC, Specificity, Sensitivity	ROC-AUC = ~0.69 ± 0.08	No	Moderate performance, no real clinical testing
Cao et al. [21]	Cardiotocography (CTG) dataset	Multimodal deep neural networks	Accuracy, Specificity, Sensitivity, AUC	Accuracy = 90.77%	No	Potential overfitting, generalization issue
Wong et al. [22]	Cardiotocography (CTG) dataset	Kolmogorov–Arnold Networks (KANs)	Accuracy, RMSE, R ²	Accuracy = 92.6%	No	Complicated design of KAN
Swain et al. [23]	A specialized dataset of 1500 fetal health records	RF, LR, GB, DT, SVM, NB, Voting	Accuracy, Precision, Recall	Accuracy = 94.02%	No	Lack of external validation
Akbulut et al. [24]	A custom clinical dataset of 96 pregnant women	SVM, LR, NN, LR, DT, DF, DJ (Decision Jungle), Bayes Point Machine	Accuracy, F1-Score, AUC	Accuracy = 89.5%	No	Small dataset size
Vani [25]	Cardiotocography (CTG) dataset	DNN, SVM	Specificity, Sensitivity, F1-Score, Gmean	Accuracy = 91%	No	Restricted model comparison, limited dataset
Chen et al. [26]	Cardiotocography (CTG) dataset	weighted random forest (WRF), BP, RF, DT, SVM, KNN	Sensitivity, Specificity, F1-score, AUC, Running time (s), Prediction error	F1-score = 97.85%	No	Limited by feature-based modeling

Table 1. Cont.

Study	Dataset	Techniques	Evaluation Metrics	Highest Performance	Explainability Techniques	Limitations
Fasihi et al. [27]	Cardiotocography (CTG) dataset	CNN	Accuracy, Sensitivity, Specificity, AUC	Accuracy = 97.46%	No	lack of external validation
Li and Liu [28]	Cardiotocography (CTG) dataset	Soft Voting, Stacking	Accuracy, Precision, Recall, F1-score, AUC, MCC	Accuracy = 95.9%	No	Inadequate data augmentation methods
Sirisha et al. [29]	Cardiotocography (CTG) dataset	LR, DT, RF, XGB, KNN, DMLP	Accuracy, Precision, Recall, F1-score	Accuracy = 97%	No	Absence of explainability techniques
Suganthi and Basilica [30]	Cardiotocography (CTG) dataset	CNN, ANN, radial basis function networks	Accuracy, MSE, RMSE, MAE, R ²	Accuracy = 96.47%	No	Limited dataset size
Salini et al. [31]	Cardiotocography (CTG) dataset	RF, DT, KNN, LR, SVM, GB	Accuracy, Precision, Recall, F1-score	Accuracy = 93%	No	Risk of overfitting, noisy, and limited data
Nagavibha et al. [32]	Cardiotocography (CTG) dataset	NB, LR, DT, RF, KNN, SVM, GB, LDA, QDA, LGBM	Accuracy, Precision, Recall, F1-score	Accuracy = 96.9%	No	Limited Generalization
Mehbodniya et al. [33]	Cardiotocography (CTG) dataset	SVM, RF, MLP, KNN	Accuracy, Precision, Recall, F1-score, Support	Accuracy = 94.5%	No	Limited real-time validation

3. Materials and Methods

The present section describes the methodological framework used in this study to construct the recommended fetal health prediction framework. It starts with an overview of the CTG dataset and its clinical characteristics, then goes over the data preprocessing processes used to maintain data quality and consistency. The next section discusses the training strategy, which includes stratified cross-validation and a leakage-free validation design, and additionally, the procedures utilized to alleviate class imbalance. Lastly, the framework and working procedure of the proposed hybrid TabNet-Residual MLP model are explained, emphasizing its significance in boosting prediction performance and resilience for fetal health classification.

3.1. Dataset

The experiments in the current study were carried out using the publicly available Cardiotocography (CTG) dataset, which is extensively used to measure fetal health [34]. This original dataset consists of 2126 samples, each representing 21 clinical parameters obtained from fetal heart rate (FHR) and uterine contraction (UC) signals, and also a target variable reflecting fetal health status.

The target variable classifies fetal complications into three clinically relevant categories:

- Class 1 (Normal): A healthy fetal state that poses no immediate concern.
- Class 2 (Suspect): Indicates uncertain trends that may demand additional monitoring.
- Class 3 (Pathological): Identifies unusual fetal abnormalities that may necessitate immediate medical intervention.

The CTG dataset consists of physiologically relevant features collected from fetal monitoring signals. Table 2 presents a summary of these characteristics, as well as the target variable.

Table 2. Overview of CTG Dataset Features.

Feature NO.	Feature	Description
1	baseline value	Fetal heart rate (beats per minute)
2	Accelerations	The rate of acceleration per second
3	fetal_movement	The total number of fetal movements per second
4	uterine_contractions	The total number of uterine contractions per second
5	light_decelerations	The rate of light deceleration per second
6	severe_decelerations	The rate of severe decelerations per second
7	prolongued_decelerations	The total number of extended decelerations per second
8	abnormal_short_term_variability	The percentage of time has unusually short-term variability
9	mean_value_of_short_term_variability	The average value of short-term variability
10	percentage_of_time_with_abnormal_long_term_variability	The proportion of time with unusually long-term variability
11	mean_value_of_long_term_variability	The average value of long-term variability
12	histogram_width	The width of the FHR histogram
13	histogram_min	Lowest FHR value
14	histogram_max	The maximum FHR value
15	histogram_number_of_peaks	The total number of histogram peaks
16	histogram_number_of_zeroes	The total number of histogram zeros
17	histogram_mode	The most frequent FHR value
18	histogram_mean	The mean FHR value
19	histogram_median	The median FHR value
20	histogram_variance	Variability of the FHR signal
21	histogram_tendency	FHR Trend Changes
22	fetal_health	Class labels indicate fetal status: Normal (0), Suspect (1), Pathological (2)

Regarding the small size of the original dataset, which may limit the generalization abilities of machine learning models, particularly deep learning approaches, a data augmentation strategy was used. This is especially essential because the proposed study intends to create and test a hybrid deep learning model, which normally requires large-scale data to efficiently understand complicated patterns while avoiding overfitting.

To overcome this restriction, synthetic samples were created using a large language model-assisted generative technique, in which the model was guided to understand the underlying statistical patterns and feature interactions observed in the original CTG dataset. The goal of this augmentation method is to allow the proposed hybrid model to fully utilize its learning capability and to conduct a systematic examination of its performance on a significantly bigger dataset.

To guarantee data consistency and reliability, the created synthetic samples were constructed to keep the original dataset's statistical features, such as feature distributions and feature interaction correlations. The augmented dataset, known as the Augmented CTG dataset, contains 50,000 samples that are both real and AI-generated. This large-scale dataset strengthens the training process and increases the model's capacity to generalize to previously undiscovered information. Moreover, to ensure computational compatibility with deep learning techniques, the class labels were converted from their original encoding (1, 2, 3) to a zero-based representation (0, 1, 2).

To guarantee compliance with real-world clinical population features, the expanded CTG dataset was built in a way that maintains the original class distribution. In particular, the percentage of synthetic samples was created to preserve the same class ratio, roughly 77% Normal, 14% Suspect, and 9% Pathological, observed in the original dataset. Instead of adding false class balancing, this design decision guarantees that the enhanced dataset reflects actual fetal health prevalence trends. By preserving the original distribution, the suggested dataset more accurately depicts the proportions of the CTG population in the actual world, improving the produced model's external credibility and clinical relevance while allowing for reliable training of the suggested hybrid architecture.

3.2. LLM-Aided Rule-Based Synthetic Data Generation

To overcome the original CTG dataset's limited size and class imbalance, a synthetic dataset of 50,000 samples was created using an LLM-supported rule-based generation technique. Claude (Anthropic) was assigned to help build a medically consistent generation technique, instead of simply acquiring knowledge or predicting data distribution. The goal was to create more synthetic samples while maintaining the clinical features, feature ranges, and class distribution identified in fetal cardiotocography (CTG) data. The main prompt given to Claude was expressed in the following manner:

"Suggest a suitable synthetic data generation approach for a medical tabular dataset with 2126 CTG records. Create a synthetic dataset of 50,000 samples while retaining the original feature schema, clinically significant value ranges, class distribution, and statistical properties. Ensure that the generated data is medically credible and appropriate for fetal health classification."

The dataset structure, which included feature names, variable kinds, and clinically significant value ranges, was identified by consulting the original CTG dataset, which had 2126 samples. The synthetic dataset was created using a predetermined class-conditional rule-based approach adapted from the FIGO (2015) medical guidelines and published CTG class incidence data reported by Ayres-de-Campos et al. (2000) [35] instead of fitting a statistical generative model to replicate the original data distribution. Claude (Anthropic) was recruited to help develop the generation process and convert the medical rules into a workflow for the structured production of synthetic data. The generation approach was

created to avoid directly replicating individual patient records while maintaining clinically realistic feature qualities.

The target fetal health class was determined prior to feature development using class-conditional sampling. Class labels were produced at random using the prevalence provided by Ayres-de-Campos et al. (2000) [35], with probabilities of 77% Normal, 14% Suspect, and 9% Pathological, keeping the class distribution consistent with actual CTG datasets as shown in Table 3.

Table 3. Proportion of target class labels in the augmented CTG dataset.

Class	Label	Class Proportion Percentage	Total Number of Each Class Instances
Normal	0	77%	38,597
Suspect	1	14%	6932
Pathological	2	9%	4471

After specifying the class label, each CTG feature was created independently using clinically approved ranges given by the FIGO 2015 recommendations. As an example, baseline fetal heart rate data for the Normal classes were drawn from a range of 110–160 bpm, whereas values for the Suspect and Pathological instances were drawn from clinically specified ranges. To establish medical validity, similar class-specific limitations were applied to the remaining CTG features.

To improve realism, Gaussian noise (1–2%) was utilized on continuous features. Quality control ensured that the generated data retained the desired class proportions, clinically relevant feature values, and descriptive statistical properties from the original CTG dataset. The synthetic dataset was created once before model training, using specified clinical guidelines and literature-based distributions. Although no individual patient records were repeated, the utilization of the original dataset as a reference for the generation technique could introduce a potential prior exposure impact, which is considered a study limitation.

3.3. Data Preprocessing

Before training the model, a structured preprocessing pipeline was used to guarantee the consistency and dependability of the input data. Initially, the standardization technique was used for feature scaling, transforming each feature to have an integer variance and a zero mean. To increase the degree of convergence and performance of machine learning and deep learning models, it is crucial to remove scale inequalities among features as defined in Equation (1):

$$Z = \frac{X - \mu}{\sigma} \quad (1)$$

where X is the original feature value, μ is the feature mean, and σ is the standard deviation. By scaling the data to have a mean value of zero and unit variance, this transformation enhances the learning algorithms' reliability and accuracy.

3.4. Training Strategy

To accomplish accurate model generalization and performance evaluation, a stratified 5-fold cross-validation technique was used. This technique maintains the original class distribution within each fold, which is especially crucial considering the CTG dataset's imbalance. By maintaining similar class proportions throughout training and validation subsets, stratification reduces sampling bias and guarantees that all classes are represented fairly during model training.

The dataset is divided into five mutually separate folds, with one fold used for validation and the remaining folds utilized for training. This procedure ensures an accurate

separation of training and validation data, thereby preventing data leakage, a significant problem in numerous fetal health prediction research studies. Model performance is provided as mean $\pm$ standard deviation across all folds, providing a statistically accurate indicator of generalizability. Furthermore, confusion matrices are constructed for each fold and then aggregated to allow for a thorough examination of class-based misclassification patterns and trends.

Figure 1 depicts stratified 5-fold cross-validation, which divides the dataset into five folds while keeping the class distribution in each fold. In each iteration, one fold is utilized for validation, and the rest for training, and the best possible performance is produced by averaging the results across all folds.

Iteration	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
1	Validation	Training	Training	Training	Training
2	Training	Validation	Training	Training	Training
3	Training	Training	Validation	Training	Training
4	Training	Training	Training	Validation	Training
5	Training	Training	Training	Training	Validation

Figure 1. Stratified 5-fold cross-validation with retained class distribution and repeated training and validation splits.

3.5. Class Imbalance Handling

To further deal with the problem of class imbalance, the proposed model includes an adaptive Focal Loss function during training. This loss function dynamically prioritizes difficult-to-classify samples while minimizing the influence of correctly classified examples. The algorithm uses class weighting factors (α), pt is the anticipated probability of the true class, and a focusing parameter (γ) controls how each sample contributes to the overall loss as defined in Equation (2):

$$FL(pt) = -\alpha t (1 - pt)^\gamma \text{Log}(pt) \quad (2)$$

Throughout this experiment, focused loss was applied to the fetal health classification problem, which included Normal, Suspect, and Pathological classes.

As defined in Equation (3), C represents the number of classes, p_c is the predicted probability for class c , and α_c is the class-specific weighting factor used to resolve class imbalance. The focusing parameter, γ , decreases the loss contribution from well-classified cases while highlighting difficult and minority-class data. This ensures that underserved classes perform better in learning.

$$FL(pt) = -\sum_{c=1}^C \alpha_c (1 - p_c)^\gamma \text{Log}(p_c) \quad (3)$$

The addition of Focal Loss improves the model's sensitivity to minority classes, especially the malignant class, which is crucial in clinical practice. The suggested approach matches actual-life clinical requirements by emphasizing challenging and underrepresented situations. Accurate diagnosis of high-risk prenatal abnormalities is critical.

3.6. Proposed Hybrid Model

The proposed methodology proposes a hybrid deep learning structure for reliable fetal health prediction using cardiotocography (CTG) data. Figure 2 provides an overview of the suggested hybrid model that uses CTG data to increase fetal health prediction accuracy by incorporating TabNet for feature learning and selection with a Residual MLP classifier. The model combines TabNet's abilities regarding interpretation-based feature representation learning with a Residual Multi-Layer Perceptron (Residual MLP) to improve non-linear classification. Firstly, TabNet is used to autonomously acquire significant feature representations by applying an attention-driven sequential decision procedure, allowing for the successful identification of clinically significant CTG features. These acquired embedded features are then transmitted into the Residual MLP, which improves and refines the representations with deep residual connections to increase gradient propagation and classification reliability. To tackle the issue of class imbalance, the entire design is optimized with Focal Loss, which focuses on difficult-to-classify samples and increases minority class sensitivity. To ensure generalizability and robustness, the model is trained from start to finish using stratified 5-fold cross-validation. Finally, performance is measured using a variety of metrics, and interpretability is ensured by utilizing SHAP and LIME-based explainability algorithms, which provide both local and global insights into model decisions.

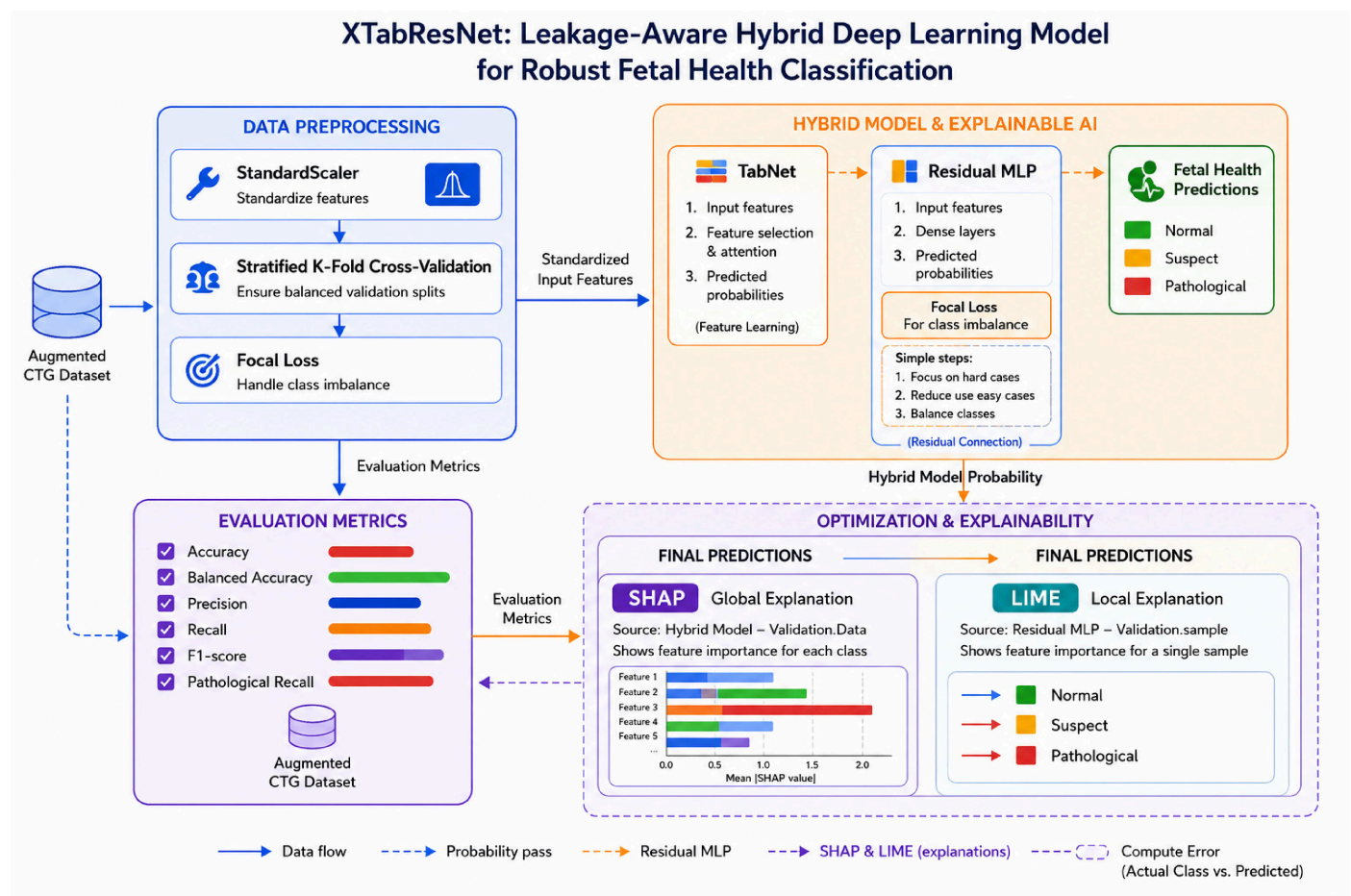


Figure 2. The Overall architecture of the Hybrid Prediction Model.

3.6.1. TabNet-Based Feature Representation Learning

After preprocessing, which includes enhanced CTG data (actual and synthetic samples), feature scaling with StandardScaler, and stratified 5-fold cross-validation, the standardized input data is passed into the TabNet architecture as the initial stage of the learning

pipeline. TabNet is a deep learning model that was intended for tabular data, making it ideal for cardiotocography (CTG) datasets with organized and heterogeneous features.

Instead of using traditional neural networks, TabNet uses a sequential attention mechanism that allows for dynamic feature selection at each decision phase. This method is implemented by attentive transformers, which construct minimal feature filters that determine whether features are highlighted or hidden during each forward pass. As a result, the model learns to assign priority to the most informative medical features, like fetal heart rate variation and uterine contraction behavior, while minimizing the impact of irrelevant or redundant data.

This framework consists of numerous decision processes, each of which refines the feature representation by incorporating existing knowledge with newly acquired information. Feature usage is also implemented utilizing global feature transformers, which improve parameter efficiency and learning consistency. This recurrent refinement approach enables TabNet to capture the complex, non-linear correlations found in prenatal health data.

The TabNet module concludes with two significant outputs: The high-level hidden feature representations contain the underlying framework of the CTG data.

The preliminary class probability distributions indicate the model's unreliable judgment of fetal health status. These outcomes serve as the starting point for the upcoming classification stage.

3.6.2. Residual MLP (Residual Multi-Layer Perceptron) Classifier

The final classification phase of the proposed hybrid architecture includes a Residual Multi-Layer Perceptron (Residual MLP), which receives TabNet's class probability outputs for further refinement. The Residual MLP is designed to train supplementary nonlinear feature representations, allowing for more discriminative fetal health categorization while keeping the predictive information retrieved by TabNet.

The Residual MLP is composed of three fully connected layers. The first layer converts the TabNet probability outputs to a 64-dimensional latent representation, followed by a ReLU activation function. The second fully connected layer preserves the 64-dimensional feature space, followed by a residual (skip) connection that transfers the input representation directly to the second layer's output. A second ReLU activation is then used before transmitting the refined features to the final fully connected layer, which maps the learnt representation to the three fetal health classes (normal, suspect, and pathological).

The residual connection makes gradient propagation more stable during training, solves the vanishing gradient problem, and saves crucial feature details across layers. By merging TabNet's feature selection capabilities with Residual MLP's nonlinear representation learning, the suggested hybrid framework increases its ability to identify between fetal health classes, especially the minority problematic class. The network is optimized with the Adam optimizer at a learning rate of 0.001 and trained for 50 epochs using Focal Loss, which improves detection of difficult and underrepresented data.

3.6.3. Focal Loss Optimization and Backpropagation

To efficiently deal with class imbalance, a prevalent feature of CTG datasets, the training procedure uses Focal Loss for the objective function. In contrast to ordinary cross-entropy loss, Focal Loss employs a modifying factor to lower the contribution of easily categorized samples while increasing the emphasis on incorrectly classified or minority-class instances. It guarantees that the model does not get biased toward the majority class while being sensitive to clinically significant but minority disorders.

Throughout training, the difference between the Residual MLP's projected likelihood distribution and the actual truth labels is calculated. The resulting loss is then transmitted

backward through the network with the backpropagation technique, allowing for gradient-driven optimization for all model parameters. The TabNet feature extraction module and the Residual MLP classifier are both updated simultaneously from start to finish, guaranteeing that feature learning and classification are fully integrated and continuously optimized.

An effective optimization technique (such as the Adam optimizer) is often employed to repeatedly update the model weights, reducing the loss function over subsequent epochs. This iterative approach is repeated throughout all folds of the stratified cross-validation architecture to ensure robust and unbiased model training.

3.6.4. Final Prediction Output

After training, the proposed hybrid system produces the final fetal health prediction using the highest probability score given by the Residual MLP classifier. The result represents the model's decision on the fetal state, which is classified into clinically important categories. The combination of TabNet and Residual MLP guarantees that the prediction is based on both readable feature selection and deep nonlinear classification, which improves overall diagnostic reliability.

In brief, the recommended approach leverages a novel two-stage hybrid architecture that incorporates the capabilities of TabNet and a Residual Multi-Layer Perceptron (Residual MLP) to improve fetal health prediction using CTG data. Throughout the first phase, TabNet is used as a feature extraction framework, where it understands complex, nonlinear relationships between input features and executes intuitive feature selection using its attention mechanism. Through the stratified 5-fold cross-validation paradigm, TabNet is trained individually in each fold, providing class probability distributions based solely on unseen validation data, assuring fair evaluation. For example, given a CTG sample, TabNet generates a probability scalar like [0.82, 0.13, 0.05], which represents the probability that the sample belongs to the Normal, Suspect, or Pathological classes, respectively. These probabilities indicate the model's trustworthiness in all classes and are subsequently passed on to the second phase.

During the second phase, a Residual MLP classifier is used as a meta-learner, with TabNet probability outputs as input. The Residual MLP has skip (residual) connections that increase gradient flow, feature reuse, and training stability. Particularly, the network learns a modification of the input probabilities while retaining the original information by including it in the converted output. This enables the model to improve its initial predictions and provide new probability values, including [0.88, 0.09, 0.03], resulting in an accurate final classification.

Focal Loss is utilized between the actual truth labels and the improved predicted probabilities to optimize the overall hybrid architecture. This effectively addresses class imbalance by highlighting samples that are difficult to categorize. Backpropagation is then used to propagate the calculated loss backward via the Residual MLP and TabNet, allowing for weight updates from start to finish for both model phases. The joint optimization of the categorization of feature components is ensured by this iterative training approach. The category with the highest likelihood is eventually selected to make the final prediction, producing a reliable and broadly applicable model for classifying fetal health.

In a simple way, the suggested hybrid framework, as shown in Figure 3, uses TabNet to process CTG features in order to produce class probability distributions. A Residual MLP classifier then refines these distributions, and Focal Loss is used to optimize the model via backpropagation to handle class imbalance and enhance prediction performance.

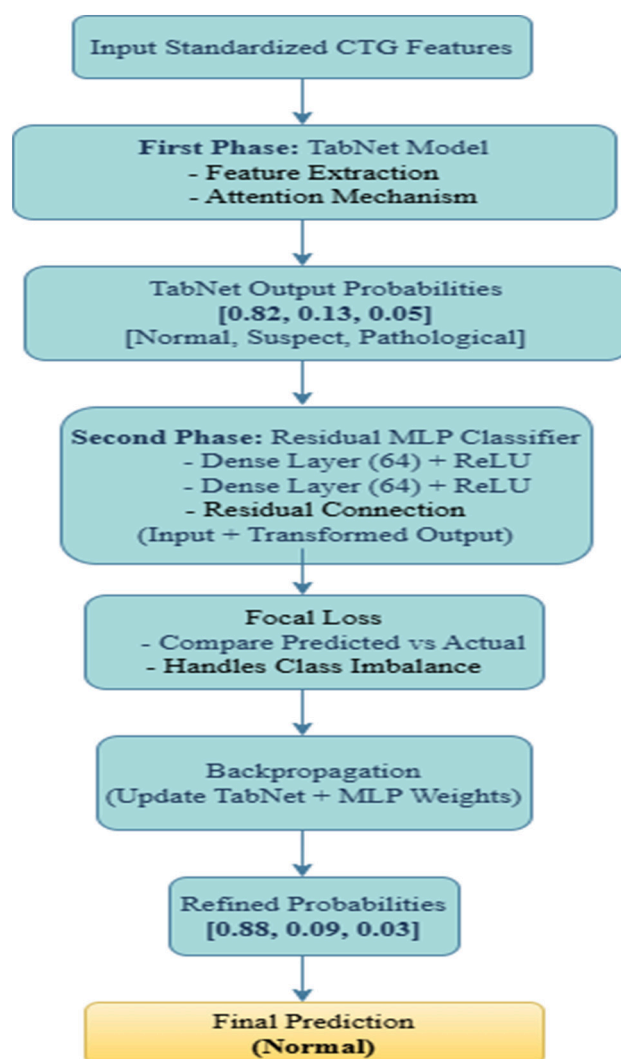


Figure 3. In-depth sequential flow of the proposed hybrid prediction model.

3.6.5. Explainability Techniques Using SHAP and LIME

Two mutually beneficial Explainable Artificial Intelligence (XAI) strategies, Shapley Additive Explanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME), were used to improve the clarity and clinical reliability of the proposed hybrid deep learning model. By identifying the characteristics that most strongly influence prediction outcomes, these techniques offer perspectives on the framework's decision-making process.

By measuring each attribute's contribution to the model's predictions over the whole dataset, SHAP was used to give global model interpretability. SHAP, which relies on cooperative game theory, gives each feature a Shapley value. This allows for a consistent evaluation of feature relevance and reveals the overall elements impacting fetal health classification.

On the other hand, local explanations for each predicted outcome were produced using LIME. LIME finds the characteristics that most strongly affect a given prediction by approximating the behavior of the complicated prediction model using an observable substitute model in nearby areas of a specific case. Analyzing clinically significant cases, such as abnormal fetal states, is a particularly useful application of this expertise.

The combination of SHAP and LIME enables both local and global interpretability, allowing for a more in-depth investigation of model behavior while increasing clarity, reliability, and clinician trust in the proposed fetal health prediction method.

3.6.6. Model Evaluation

To guarantee a thorough and trustworthy evaluation under the class imbalance conditions present in CTG datasets, the performance of the indicated hybrid model is measured using a collection of selective classification measures. In particular, many characteristics of model performance are captured using Accuracy, Balanced Accuracy, Precision, Recall, Recall (Pathological), and F1-score as depicted in Table 4. Accuracy may be weighted toward the majority class even if it offers a general indicator of correctly classified samples. As a result, Balanced Accuracy is also considered since it provides a more representative assessment of the model's performance by considering class imbalance by averaging recall among all classes.

Table 4. Evaluation Measures Utilized for the Proposed Hybrid Model.

Evaluation Measure	Equation
Accuracy	$\frac{TP+TN}{TP+TN+FP+FN}$
Balanced Accuracy	$\frac{1}{C} \sum_{i=1}^C Recall_i$ C represents the number of classes, which is 3
Precision	$\frac{TP}{TP+FP}$
Recall	$\frac{TP}{TP+FN}$
$Recall_{Pathological}$	$\frac{TP_{Pathological}}{TP_{Pathological}+FN_{Pathological}}$
F1_Score	$2 \times \frac{Precision * Recall}{Precision + Recall}$

TP (True Positive): The real class is Pathological, and the model accurately predicts Pathological. TN (True Negative): The real class is not Pathological, and the model correctly predicts it as not Pathological. FP (False Positive): The real class is not Pathological, but the model incorrectly predicts it as Pathological. FN (False Negative): The real class is Pathological, but the model incorrectly predicts it as Normal or Suspect. $TP_{Pathological}$: The real class is Pathological, and the model correctly predicts Pathological. $FN_{Pathological}$: The real class is Pathological, but the model incorrectly predicts Normal or Suspect.

The model's capability to find suitable examples is evaluated using precision and recall. While recall examines the model's capacity to accurately identify real positive instances, precision calculates the percentage of properly estimated positive cases among all predicted positives. Because of its clinical significance, Recall (Pathological) is given special attention in this case. The pathological class includes high-risk fetal disorders that need to be treated right away; if these instances are not accurately identified, there could be significant consequences. Even when there is a class imbalance, the model favors avoiding false negatives in this crucial class by emphasizing recall (pathological). Because of this, it is a crucial measure of the model's practical dependability in real-life healthcare settings.

Another balanced evaluation that takes into consideration both false positives and false negatives is provided by the F1-score, which is calculated from the harmonic means of precision and recall. Because it illustrates the trade-off between sensitivity and precision, this metric is especially helpful in situations where the distribution of classes is not equal.

Overall evaluation indicators are calculated using a stratified 5-fold cross-validation design, with each fold preserving the original class distribution. The framework is trained and validated independently across all folds, with the final results provided as the fold-average performance. This assessment technique increases the proposed model's resilience and predictive capability while reducing bias and overfitting.

In addition to performance measurement, model interpretability is an important component of the proposed approach, especially given the physiological nature of cardiotocography (CTG) data, where understanding the reasons behind predictions is key for making medical decisions. To solve this issue, the suggested model combines SHAP (SHapley Additive Explanations) and LIME (Local Interpretable Model-agnostic Explanations)

as complementary explainability strategies. SHAP is used to give global interpretability by calculating the contribution of every input attribute to the overall model predictions using Shapley values. This makes it possible to identify the most important physiological characteristics in the setting of CTG analysis, such as fetal heart rate baseline, variability, accelerations, and uterine contractions, which are crucial in establishing the health state of the fetus. Professionals can align model behavior with accepted medical knowledge by using SHAP overview plots and feature importance rankings to determine which features consistently influence the model's judgments over the whole dataset.

On the other hand, LIME is used to explain specific predictions to give local interpretability. LIME uses a more straightforward interpretable model in the neighborhood of a particular CTG record to estimate the performance of the sophisticated hybrid model. This makes it possible to determine which characteristic values, like anomalous variability or decreased accelerations, favorably or unfavorably affect a specific classification outcome (like Pathological or Suspect). In clinical settings, when comprehending the logic underlying a single risky prediction is essential for enhancing healthcare decisions, similar instance-based explanations are very helpful.

The combination of SHAP and LIME guarantees a thorough explainability framework that encompasses both local selection behavior and global patterns. Through offering comprehensible and clinically significant insights into fetal health prediction, this multi-level interpretability not only increases awareness and confidence in the suggested hybrid model but also makes it easier for it to be implemented in actual clinical settings.

4. Results

4.1. Proposed Hybrid Model Performance Results

The efficiency of the proposed hybrid TabNet-Residual MLP model was measured using the outlined classification criteria within the stratified cross-validation framework. The framework reveals good predictive capacity across all evaluation criteria, supporting the suggested architectural framework's efficacy in CTG-based fetal health classification. As demonstrated in Table 5, the model attained an accuracy of 99.30% and a balanced accuracy of 99.10%, suggesting strong performance even under class imbalance conditions. The model also achieved a recall of 99.10% and a precision of 98.39%, demonstrating its capacity to accurately identify appropriate instances while preserving low misclassification rates.

Table 5. Results of the Proposed Hybrid Prediction Model.

Evaluation Metric	Value
Accuracy	99.30%
Balanced Accuracy	99.10%
Precision	98.39%
Recall	99.10%
<i>Recall</i> _{Pathological}	98.65%
F1_Score	98.73%

Specifically, the suggested model acquired a Recall (Pathological) of 98.65%, suggesting its remarkable capacity to predict high-risk fetal diseases. This is an important finding since reducing missing pathological cases is critical in clinical settings. Furthermore, the F1-score of 98.73% demonstrates an appropriate trade-off between precision and recall, verifying the model's overall reliability. These findings support the efficacy of combining TabNet with a Residual MLP and Focal Loss, resulting in accurate and strong fetal health prediction. Also, the suggested model's high Recall (Pathological) demonstrates its medical reliability, as it reduces the chance of missing significant fetal health issues.

The detected performance gain cannot be attributed exclusively to the increased quantity of synthetic samples. In fact, it indicates the cumulative impact of the proposed TabNet-Residual MLP architecture, Focal Loss for class imbalance resolution, medically limited classes-conditional synthetic data creation, and increased training data diversity. While the synthetic data were created using knowledge from the original dataset and known clinical recommendations, their primary purpose is to increase the model's generalization rather than offer wholly new information. As a result, the claimed performance should be understood in terms of these integrated methodological advancements.

The proposed framework outperformed the performance reported by the majority of earlier CTG-based fetal health prediction studies (usually ranging from 93% to 97%), as indicated in Table 5, with an accuracy of almost 99%. Several complementary elements are responsible for this improvement. First, the model was able to learn more representative patterns than those found in the initial limited CTG dataset because the LLM-aided rule-based synthetic data generation framework generated a huge, medically consistent dataset that retained feature relationships and class-specific characteristics. Second, while Focal Loss enhanced the identification of minority abnormal cases, the hybrid TabNet-Residual MLP architecture successfully coupled TabNet's variable feature selection with the Residual MLP's nonlinear representation learning capacity. Lastly, a thorough evaluation of the suggested framework was made possible by the application of stratified five-fold cross-validation. When taken as a whole, these methodological improvements helped the suggested strategy attain better prediction performance.

4.2. Comparative Analysis of the Proposed Model with Existing Methodologies and Mitigating Data Leakage

Table 6 compares the proposed hybrid model to existing ensemble-based techniques across different data handling techniques. The original ensemble model indicates great performance (98.9% accuracy); however, this result was acquired using a pipeline with data leakage, in which augmentation is applied before data splitting, resulting in overstated performance.

Table 6. Evaluation of baseline and proposed hybrid model, emphasizing the importance of data leakage mitigation and enhanced pathological prediction.

Model	Baseline/Improved	Dataset	Data Handling	Accuracy	Balanced Accuracy	Precision	Recall	Recall (Pathological)	F1_Score
Ensemble Model (Prior Study) [17]	Baseline	Original Dataset	With Leakage	0.989	-	0.990	0.986	-	0.993
Ensemble Model (Prior Study) [17]	Baseline	Augmented Dataset	With Leakage	0.9680	0.9790	0.9733	0.9680	0.9836	0.9692
Ensemble Model (Prior Study) [17]	Baseline	Augmented Dataset	Without Leakage	0.9314	0.7507	0.9361	0.9314	0.3478	0.9194
Proposed Model (TabNet + ResMLP + Focal Loss)	Full Model	Original Dataset	Original CTG Data	0.8627	0.7620	0.7733	0.7620	0.6759	0.7610
Proposed Model (TabNet + ResMLP + Focal Loss)	Full Model	Augmented Dataset	Synthetic + Original CTG Data	0.9930	0.9910	0.9839	0.9910	0.9865	0.9873

After adjusting the pipeline by separating data prior to augmentation, the model's performance drops significantly to 93.14% accuracy and 75.07% balanced accuracy, indicating a more realistic evaluation. This reveals that the previously stated improvements were

mostly impacted by inappropriate data processing. The effects of data leakage are further measured by a performance gap in accuracy and balanced accuracy between leaking-prone and leaking-free environments.

However, when compared to an effectively built, leakage-free pipeline, the suggested solo model outperforms all others, as displayed in Table 6.

4.3. Explainability Analysis Results

4.3.1. Explainability Analysis Using SHAP

To strengthen the understanding of the suggested hybrid model's decision-making behavior, SHAP (Shapley Additive Explanations) was utilized to present global feature importance across cross-validation folds. SHAP, as depicted in Figure 4 and quantitatively summarized in Figure 5, reveals the most significant CTG attributes that contribute to fetal health classification. The findings show that prolonged decelerations are the most prevalent feature, with the highest mean SHAP value, implying that they play the most important role in identifying fetal health states, specifically risky pathological cases. This observation is biologically significant since prolonged decelerations are significantly linked to fetal pain.

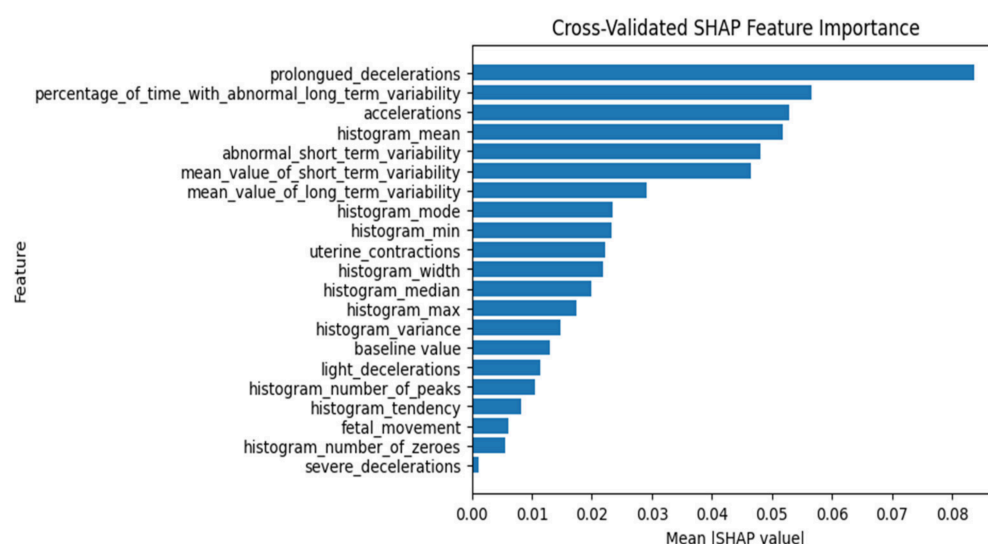


Figure 4. Cross-validated SHAP feature importance demonstrates the most significant CTG features influencing the predictions derived from the suggested model.

Furthermore, the proportion of time with abnormal long-term variability, accelerations, and histogram_mean comprises some of the top contributing variables, demonstrating that fetal heart rate variability and acceleration patterns have a considerable influence on model predictions. Short-term and long-term variability characteristics, including abnormal short-term variability, mean value of short-term variability, and mean value of long-term variability, are also significant, highlighting the importance of variability-based indicators in CTG evaluations.

Attributes like severe decelerations, histogram number of zeroes, and fetal movement, on the other hand, have relatively minor SHAP contributions, implying a smaller overall impact on the model's global predictions.

In general, the SHAP analysis shown in Figures 4 and 5 verifies that the suggested model mainly depends on clinically significant CTG features, especially variability measures and deceleration patterns, which align the model's predictive behavior with accepted fetal monitoring principles. This improves the suggested framework's clarity and promotes trust in its medical utilization.

Cross-Validated SHAP Feature Importance:

	Feature	Importance
6	prolongued_decelerations	0.083645
9	percentage_of_time_with_abnormal_long_term_var...	0.056508
1	accelerations	0.052877
17	histogram_mean	0.051863
7	abnormal_short_term_variability	0.048127
8	mean_value_of_short_term_variability	0.046420
10	mean_value_of_long_term_variability	0.029121
16	histogram_mode	0.023531
12	histogram_min	0.023238
3	uterine_contractions	0.022130
11	histogram_width	0.021909
18	histogram_median	0.019939
13	histogram_max	0.017522
19	histogram_variance	0.014807
0	baseline value	0.013062
4	light_decelerations	0.011348
14	histogram_number_of_peaks	0.010529
20	histogram_tendency	0.008194
2	fetal_movement	0.006020
15	histogram_number_of_zeroes	0.005591
5	severe_decelerations	0.001103

Figure 5. A numerical hierarchy of mean SHAP values for CTG features over cross-validation folds.

4.3.2. Explainability Analysis Using LIME

To supplement SHAP's global interpretability, LIME (Local Interpretable Model-agnostic Explanations) was utilized to examine individual predictions in standard CTG data from the Normal, Suspect, and Pathological classes. As shown in Figures 6–11, LIME gives instance-level explanations by highlighting the individual features that positively or adversely impacted each classification result.

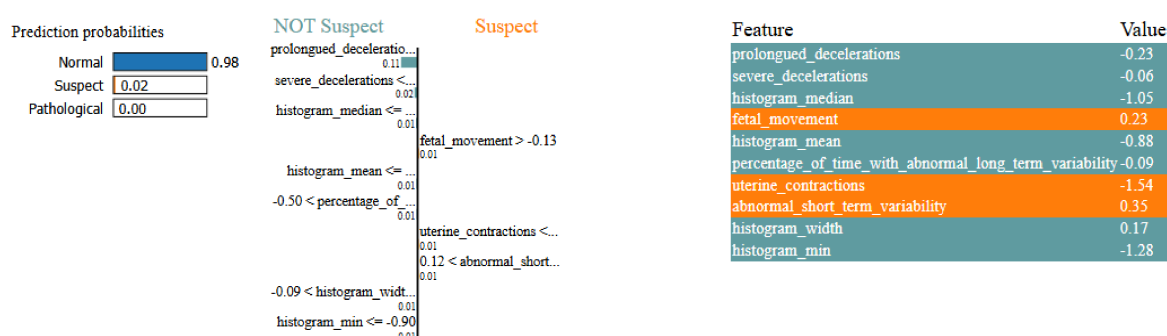


Figure 6. Normal example prediction, likelihood, and feature contribution hierarchy.

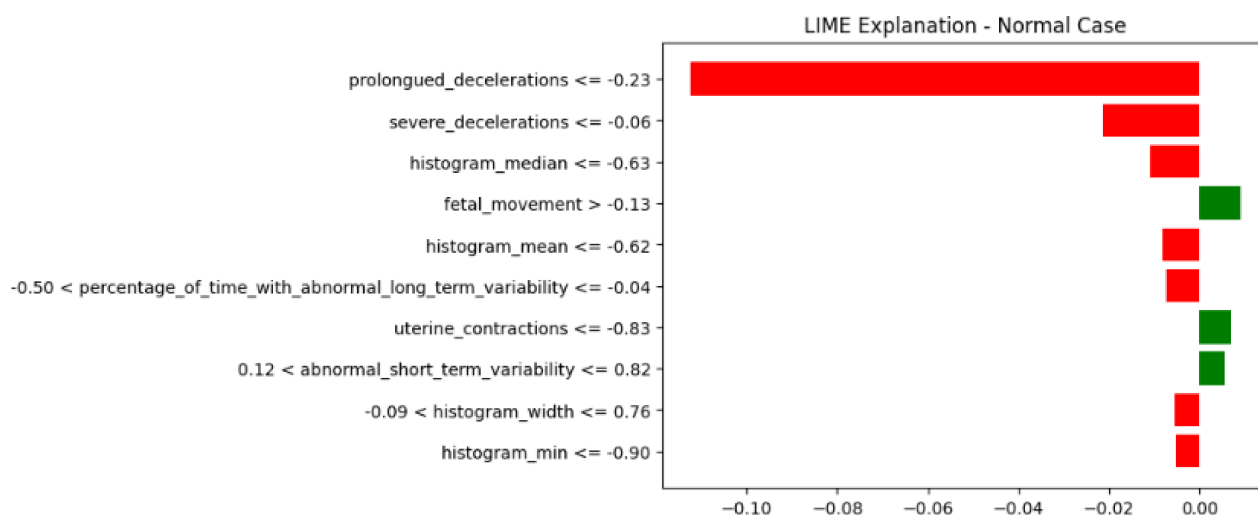


Figure 7. LIME local feature importance for the Normal example.

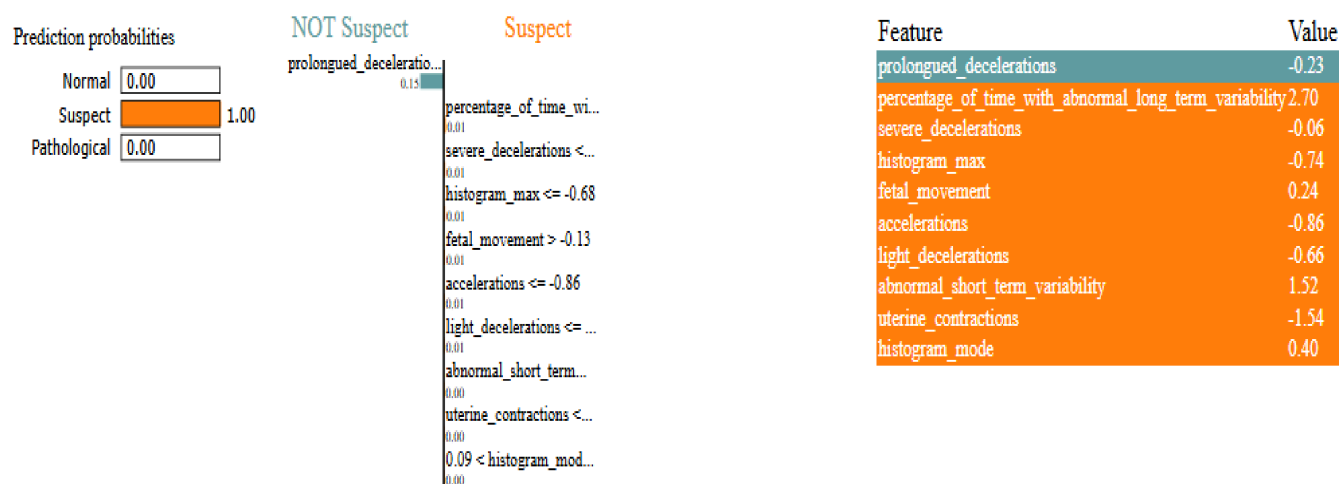


Figure 8. Suspect example prediction likelihood and feature contribution hierarchy.

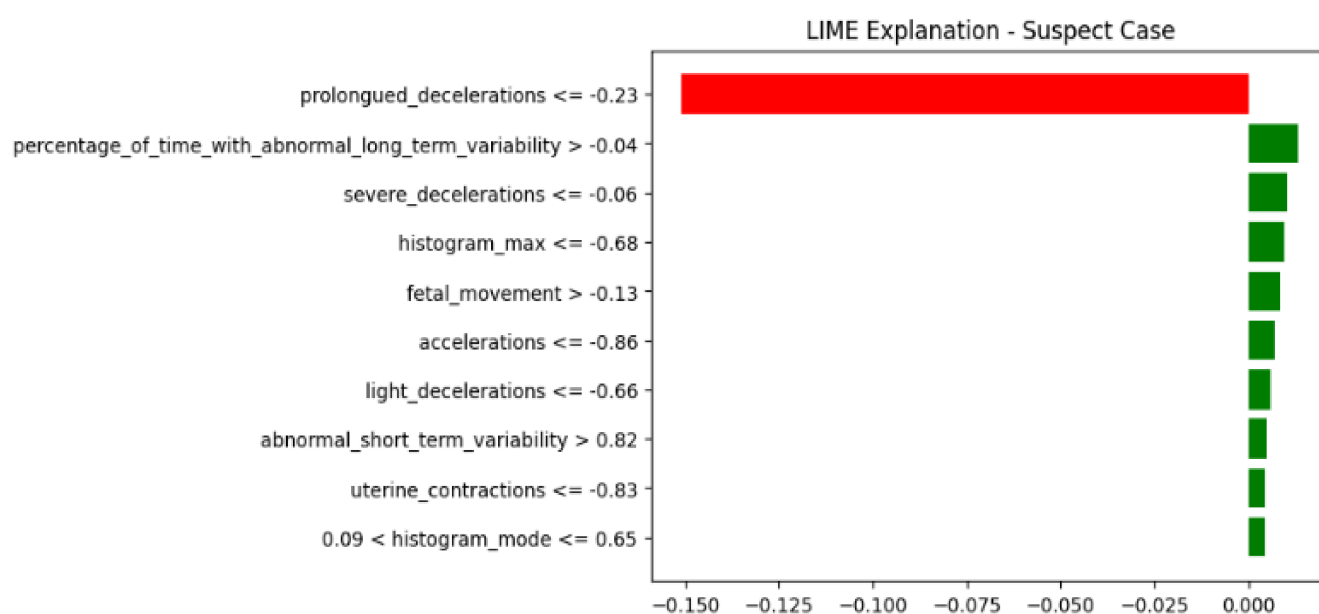


Figure 9. LIME local feature importance for the Suspect example.

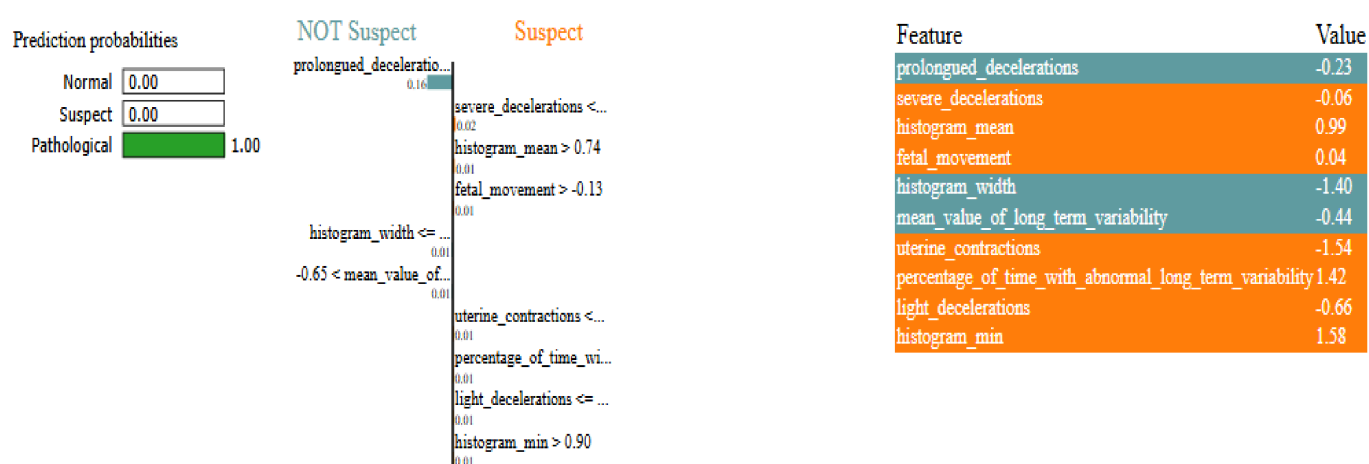


Figure 10. Pathological example prediction likelihood and feature contribution hierarchy.

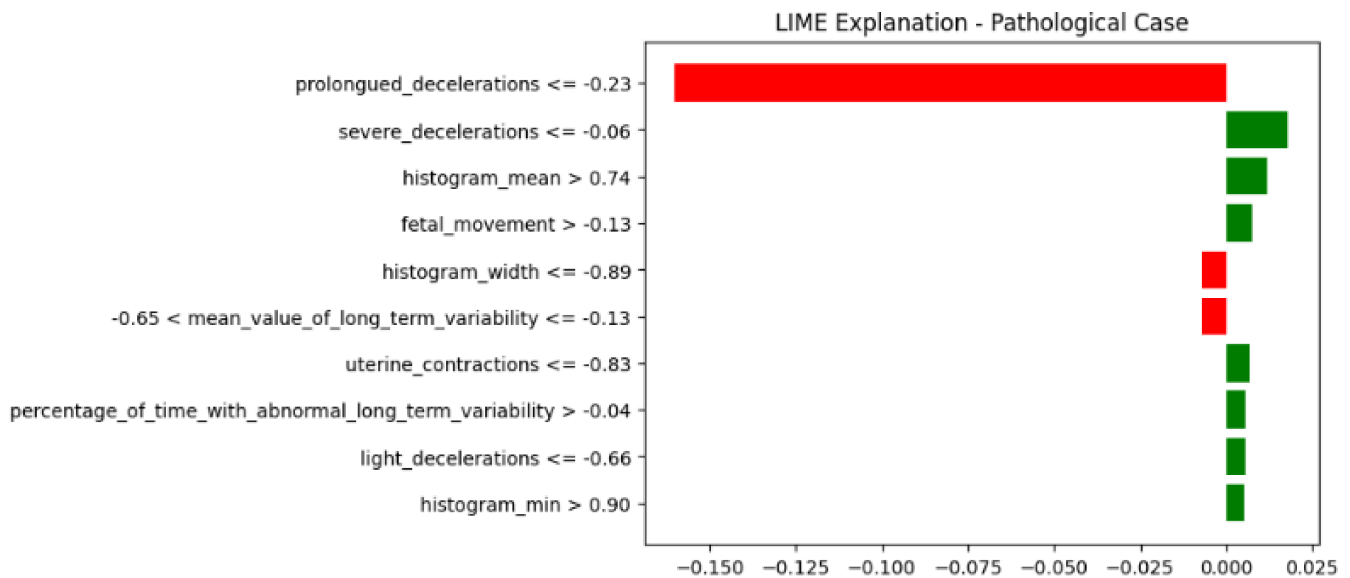


Figure 11. LIME local feature importance for the Pathological example.

Considering the Normal scenario depicted in Figures 6 and 7, the model predicts the Normal class with high confidence, owing mostly to favorable values in variables such as prolonged decelerations, severe decelerations, and histogram median, all of which contribute significantly to the Normal prediction. At the same time, a small number of attributes have tiny opposing impacts, but their combined influence is inadequate to change the overall classification. This suggests that the model recognizes uniform CTG features linked to healthy fetal situations.

The prediction for the Suspect example in Figures 8 and 9 is molded with an improved equal mixture of contributing factors, indicating the class's intermediate nature. Prolonged decelerations, proportion of time with anomalous long-term variability, and histogram variance all have a larger role in changing the prediction to Suspect, indicating moderate departures from normal fetal physiology. This highlights the model's sensitivity to modest physiological anomalies, which may indicate potential worry but lack noticeable clinical significance.

With the Pathological instance in Figures 10 and 11, LIME emphasizes that abnormal CTG indicators, particularly extended decelerations, abnormal short-term variability, and histogram-related abnormalities, greatly influence the Pathological class prediction. These attributes are clinically connected with fetal distress, and their dominant role demonstrates that the model correctly prioritizes risky physiological patterns when identifying important instances.

In general, as demonstrated in Figures 6–11, LIME analysis shows that the suggested hybrid model bases its predictions on clinically significant CTG feature patterns at the specific sampling level. LIME improves honesty, increases confidence in model predictions, and offers useful case-based interpretability for practical fetal health evaluation by demonstrating how particular features affect each case differently.

5. Conclusions and Future Work

This research study suggested a novel hybrid fetal health prediction methodology based on a TabNet-Residual Multi-Layer Perceptron (Residual MLP) architecture complemented with Focal Loss to enhance a CTG-based classification of Normal, Suspect, and Pathological fetal states. The suggested framework employs TabNet as an attentive feature learning technique capable of capturing complicated feature connections, while the Residual MLP refines probability outputs to enhance stability and prediction accuracy.

While trained on the proposed augmented CTG dataset, the hybrid framework shows remarkable standalone performance, with 99.30% accuracy, 99.10% balanced accuracy, and 98.65% recall for pathological cases, indicating strong expertise in detecting clinically significant situations.

By combining the suggested hybrid model with an already-existing ensemble framework, the study presents a universal enhancement strategy that goes beyond standalone performance and improves predictability within the original (leakage-prone) pipeline, with 95.40% accuracy and 95.39% balanced accuracy. A significant contribution of this work is the identification and mitigation of conventional train–test data leakage through an improved experimental protocol that performs data partitioning before model training and evaluation. The evaluation showed significantly lower but more realistic performance values of 78.57% accuracy and 70.00% balanced accuracy by reorganizing the pipeline to guarantee correct data splitting before augmentation, demonstrating that earlier inflated findings were primarily caused by leakage effects.

The impact of incorrect data processing is quantified by comparing a leakage-prone pipeline with the revised experimental protocol designed to prevent conventional train–test data leakage of roughly +21.8% in accuracy and +25.39% in balanced accuracy. This emphasizes how crucial it is to have a robust experimental setup in healthcare machine learning applications because methodological errors can severely bias the efficacy of the model. Significantly, using the expanded dataset, the proposed hybrid framework achieved 99.30% accuracy and 99.10% balanced accuracy, outperforming all baseline and corrected configurations. These findings show consistent predictive performance under the suggested data augmentation protocol, which prevents information leakage during model development and evaluation by generating synthetic data only after the original dataset has been split into training and testing subsets. The results demonstrate the suggested methodology's potential for precise CTG-based fetal health prediction.

The model demonstrates a notable improvement in identifying high-risk fetal abnormalities in terms of clinically relevant outcomes; pathological recall increased from 63.64% in the corrected baseline to 98.65% in the suggested hybrid model, which represents an improvement of almost +35%. This substantial enhancement shows the model's greater sensitivity for detecting problematic fetal states and illustrates its capacity to offer earlier AI-assisted fetal health evaluation, subject to further testing using independent medical datasets.

Additionally, by providing both local and global explanations of predictions, the combined use of SHAP and LIME improves model interpretability. While LIME provides particular interpretability across all fetus classes, SHAP analysis finds clinically significant CTG attributes, such as prolonged decelerations and variability-related indicators, as important factors in decision-making. These explanation findings enhance model clarity and may facilitate physician interpretation, hence supporting the future creation of credible AI-powered medical decision-support systems following extensive clinical validation.

Overall, the proposed hybrid framework performed well in the experimental conditions while effectively addressing traditional train–test data leakage, enhancing minority-class detection, and delivering interpretable predictions via explainable AI techniques. These results suggest that the proposed model is a potential approach to AI-assisted prenatal health prediction. However, further validation using alternative medical datasets and predictive clinical investigations is necessary before considering broad clinical implementation.

Despite the positive findings, this study has a number of drawbacks. First, while the proposed LLM-aided rule-based synthetic data generation framework includes medically defined limitations and multiple quality monitoring procedures, synthetic data is unable to fully represent the complexity, variability, and unique clinical characteristics found in

real-world patient populations. As a result, the model's results on synthetic data may not accurately represent how it operates when applied to real clinical datasets. Second, while standard train–test data leakage was avoided throughout model building and evaluation, the synthetic dataset was created once with the entire original dataset as a reference. As a result, the possible impact of prior exposure cannot be ruled out. Finally, separate clinical datasets from various healthcare organizations were used to evaluate the proposed framework utilizing stratified five-fold cross-validation without external validation. As a result, the applicability of the suggested framework across varied patient demographics and therapeutic situations has yet to be proven.

Future work will focus on testing the proposed framework with large-scale, multidisciplinary clinical CTG datasets obtained from various healthcare organizations to determine its applicability and clinical usefulness. Integrating periodic deep learning frameworks, like Transformer-powered or sequence-aware models, could further enhance continuous fetal health tracking. Future research will also explore potential clinical evaluation and integration into AI-driven medical decision-support systems, in collaboration with maternal experts, to assess the model's practical value in real-world healthcare settings. Future studies will also include a comprehensive computational performance analysis, including training time, inference time, computational complexity, and hardware requirements, to further examine the efficiency and adaptability of the proposed framework for real-life implementation.

Author Contributions: Conceptualization, S.A.; methodology, S.A.; software, S.A.; validation, S.A., D.S. and A.Y.; formal analysis, S.A.; investigation, S.A.; resources, A.Y.; data curation, S.A.; writing—original draft preparation, S.A.; review and editing, D.S. and A.Y.; visualization, S.A.; supervision, D.S. and A.Y.; project administration, D.S. and A.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The original Fetal Health Classification dataset is publicly available and freely accessible through the Kaggle repository (<https://www.kaggle.com/datasets/andrewmvd/fetal-health-classification>) (accessed on 27 July 2026) and the UCI Machine Learning Repository (Cardiotocography—UCI Machine Learning Repository). To validate the suggested methodology, a much bigger synthetic fetal health dataset was created from the original dataset using a generative artificial intelligence (AI) approach. The synthetic dataset was developed to improve the volume and diversity of the training data while maintaining the properties of the original dataset, and it is publicly available at: (<https://drive.google.com/drive/folders/11vOEogTFubfFpCAXOV76S9tlx5EqLVC9?usp=sharing>) (accessed on 27 July 2026).

Acknowledgments: I sincerely wish to thank Doaa and Ahmed for their support, suggestions, and insightful comments throughout this research. I truly value their consistent support and advice, and it is an immense pleasure to have learned from their knowledge, commitment, and friendliness.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

CTG	Cardiography
Residual MLP	Residual Multi-Layer Perceptron
XAI	Explainable Artificial Intelligence
AI	Artificial intelligence
ML	Machine Learning
DL	Deep Learning

DT	Decision Tree
SVM	Support Vector Machine
RF	Random Forest
SHAP	Shapley Additive Explanations
LIME	Local Interpretable Model-agnostic Explanations
PCA	Principal Component Analysis
SMOTE	Synthetic Minority Oversampling Technique
ANN	Artificial Neural Network
AUC	Area Under Curve
CNN	Convolutional Neural Network
LGBM	Light Gradient Boosting Machine
KANs	Kolmogorov–Arnold Networks
LR	Logistic Regression
NB	Naïve Bayes
AUC-ROC	Area Under the Receiver Operating Characteristic Curve
WRF	weighted random forest
MCC	Matthews Correlation Coefficient
DMLP	Dynamic Multi-Layer Perceptron
RBFNN	Radial Basis Function Neural Network
ET	Extra Tress
RNN	Recurrent Neural Network
LSTM	Long Short-Term Memory
XBG	Extreme Gradient Boosting
NN	Neural Network
DJ	Decision Jungle
BP	Backpropagation
MSE	Mean Squared Error
RMSE	Root Mean Squared Error
MAE	Mean Absolute Error
R2	Coefficient of Determination
LDA	Linear Discriminant Analysis
QDA	Quadratic Discriminant Analysis
CatBoost	Categorical Boosting
ADA	Adaptive Boosting
Kappa	Cohen’s Kappa Coefficient
FHR	Fetal Heart Rate
UC	Uterine Contraction
LLM	Large Language Model

References

1. World Health Organization. Stillbirth. Available online: <https://www.who.int/health-topics/stillbirth> (accessed on 1 July 2026).
2. World Health Organization. Global Estimates of Neonatal and Perinatal Mortality. Available online: <https://iris.who.int/server/api/core/bitstreams/481d0648-8020-4f8e-95d2-8b5bfff632f2/content> (accessed on 27 July 2026).
3. Jha, S.P.; Kumari, S.; Sinha, A.; Roy, R. Admission cardiotocography in term pregnancies for predicting intrapartum interventions and early neonatal outcomes: A systematic review. *Cureus* **2026**, *18*, e109070. [CrossRef] [PubMed]
4. American College of Obstetricians and Gynecologists. ACOG Practice Bulletin No. 106: Intrapartum fetal heart rate monitoring: Nomenclature, interpretation, and general management principles. *Obstet. Gynecol.* **2009**, *114*, 192–202. [CrossRef] [PubMed]
5. Ayres-de-Campos, D.; Spong, C.Y.; Chandrachan, E.; FIGO Intrapartum Fetal Monitoring Expert Consensus Panel. FIGO consensus guidelines on intrapartum fetal monitoring: Cardiotocography. *Int. J. Gynecol. Obstet.* **2015**, *131*, 13–24. [CrossRef] [PubMed]
6. Pereira, S.; Chandrachan, E.; Mappa, I.; Perez-Bonfils, A.G. Implementation of physiological interpretation of fetal heart rate changes: From scientific principles to frontline practice. *J. Matern. Fetal Neonatal Med.* **2026**, *39*, 2612451. [CrossRef] [PubMed]
7. Alfievic, Z.; Devane, D.; Gyte, G.M.; Cuthbert, A. Continuous cardiotocography (CTG) as a form of electronic fetal monitoring (EFM) for fetal assessment during labour. *Cochrane Database Syst. Rev.* **2017**, *5*, CD006066. [CrossRef] [PubMed]

8. Ben M'Barek, I.; Jauvion, G.; Ceccaldi, P.-F. Computerized cardiotocography analysis during labor: A state-of-the-art review. *Acta Obstet. Gynecol. Scand.* **2023**, *102*, 130–137. [[CrossRef](#)] [[PubMed](#)]
9. Le Vance, J.; Adeoye, A.; Man, R.; Eltaweel, N.; Gurney, L.; Morris, R.K.; Hodgetts Morton, V. Remote home cardiotocography: A systematic review and meta-analysis. *PLoS Digit. Health* **2026**, *5*, e0001184. [[CrossRef](#)] [[PubMed](#)]
10. Nazli, I.; Korbeko, E.; Dogru, S.; Kugu, E.; Sahingoz, O.K. Early detection of fetal health conditions using machine learning for classifying imbalanced cardiotocographic data. *Diagnostics* **2025**, *15*, 1250. [[CrossRef](#)] [[PubMed](#)]
11. Kosuru, V.V.; Ramaraju, S.V.; Rajan, A.; Santhosh, G.; Anjali, T. CTG-based fetal health prediction: A comparative study of machine learning models. In Proceedings of the 2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT), Kamand, India, 24–28 June 2024; pp. 1–7. [[CrossRef](#)]
12. Miyata, K.; Shibata, C.; Fukunishi, H.; Hemmi, K.; Kinoshita, H.; Hirakawa, T.; Urushiyama, D.; Kurakazu, M.; Yotsumoto, F. Cardiotocography-based experimental comparison of artificial intelligence and human judgment in assessing fetal asphyxia during delivery. *Cureus* **2025**, *17*, e78282. [[CrossRef](#)] [[PubMed](#)]
13. Park, C.E.; Choi, B.; Park, R.W.; Kwak, D.W.; Ko, H.S.; Seong, W.J.; Cha, H.-H.; Kim, H.M.; Lee, J.; Seol, H.-J.; et al. Automated interpretation of cardiotocography using deep learning in a nationwide multicenter study. *Sci. Rep.* **2025**, *15*, 19617. [[CrossRef](#)] [[PubMed](#)]
14. Yin, W.; Shen, Z.; Cheng, Z.; Feng, C.; Sun, G. Fetal health state detection method based on parameters efficient ensembling of deep learning. *Front. Public Health* **2026**, *14*, 1808284. [[CrossRef](#)] [[PubMed](#)]
15. Altalhan, M.; Algarni, A.; Turki-Hadj Alouane, M. Imbalanced data problem in machine learning: A review. *IEEE Access* **2025**, *13*, 13686–13699. [[CrossRef](#)]
16. Nieto, N.; Eickhoff, S.B.; Jung, C.; Reuter, M.; Diers, K.; Kelm, M.; Lichtenberg, A.; Raimondo, F.; Patil, K.R. Impact of leakage on data harmonization in machine learning pipelines in class imbalance across sites. *Neurocomputing* **2026**, *680*, 133146. [[CrossRef](#)]
17. Ahmed, S.S.; Mahmoud, N.M. Early detection of fetal health status based on cardiotocography using artificial intelligence. *Neural Comput. Appl.* **2025**, *37*, 16753–16779. [[CrossRef](#)]
18. Feng, J.; Liang, J.; Qiang, Z.; Hao, Y.; Li, X.; Li, L.; Chen, Q.; Liu, G.; Wei, H. A hybrid stacked ensemble and Kernel SHAP-based model for intelligent cardiotocography classification and interpretability. *BMC Med. Inform. Decis. Mak.* **2023**, *23*, 273. [[CrossRef](#)] [[PubMed](#)]
19. Barnova, K.; Martinek, R.; Vilimkova Kahankova, R.; Jaros, R.; Snasel, V.; Mirjalili, S. Artificial intelligence and machine learning in electronic fetal monitoring. *Arch. Comput. Methods Eng.* **2024**, *31*, 2557–2588. [[CrossRef](#)]
20. Chiou, N.; Young-Lin, N.; Kelly, C.; Cattiau, J.; Tiyasirichokchai, T.; Diack, A.; Koyejo, S.; Heller, K.; Asiedu, M. Development and evaluation of deep learning models for cardiotocography interpretation. *npj Womens Health* **2025**, *3*, 21. [[CrossRef](#)]
21. Cao, Z.; Wang, G.; Xu, L.; Li, C.; Hao, Y.; Chen, Q.; Li, X.; Liu, G.; Wei, H. Intelligent antepartum fetal monitoring via deep learning and fusion of cardiotocographic signals and clinical data. *Health Inf. Sci. Syst.* **2023**, *11*, 16. [[CrossRef](#)] [[PubMed](#)]
22. Wong, W.K.; Juwono, F.H.; Apriono, C.; Fitri, I.R. Fetal health prediction from cardiotocography recordings using Kolmogorov–Arnold networks. *IEEE Open J. Eng. Med. Biol.* **2025**, *6*, 345–351. [[CrossRef](#)] [[PubMed](#)]
23. Swain, S.; Patra, K.J.; Jayasingh, S.K.; Mohanty, M.N.; Pattanayak, B.K. The role of ensemble learning in advancing child fetal health: A focus on voting and bagging techniques. In Proceedings of the 2024 1st International Conference on Cognitive, Green and Ubiquitous Computing (IC-CGU), Bhubaneswar, India, 1–2 March 2024. [[CrossRef](#)]
24. Akbulut, A.; Ertugrul, E.; Topcu, V. Fetal health status prediction based on maternal clinical history using machine learning techniques. *Comput. Methods Programs Biomed.* **2018**, *163*, 87–100. [[CrossRef](#)] [[PubMed](#)]
25. Vani, R. Weighted deep neural network based clinical decision support system for the determination of fetal health. *Int. J. Recent Technol. Eng.* **2019**, *8*, 8564–8569. [[CrossRef](#)]
26. Chen, J.; Liu, X.-C.; Wei, H.; Chen, Q.-Q.; Hong, J.-M.; Li, Q.-N.; Hao, Z.-F. *Imbalanced Cardiotocography Multi-Classification for Antenatal Fetal Monitoring Using Weighted Random Forest*; Lecture Notes in Computer Science; Springer: Cham, Switzerland, 2019; pp. 75–85. [[CrossRef](#)]
27. Fasihi, M.; Nadimi-Shahraki, M.H.; Jannesari, A. A shallow 1-D convolution neural network for fetal state assessment based on cardiotocogram. *SN Comput. Sci.* **2021**, *2*, 4. [[CrossRef](#)]
28. Li, J.; Liu, X. Fetal health classification based on machine learning. In Proceedings of the 2021 IEEE 2nd International Conference on Big Data, Artificial Intelligence and Internet of Things Engineering (ICBAIE), Nanchang, China, 26–28 March 2021. Available online: <https://ieeexplore.ieee.org/abstract/document/9389902> (accessed on 27 July 2026).
29. Sirisha, U.; Srinivasu, P.N.; Padmavathi, P.; Kim, S.; Pavate, A.; Shafi, J.; Ijaz, M.F. Dynamic multi-layer perceptron for fetal health classification using cardiotocography data. *Comput. Mater. Contin.* **2024**, *80*, 2301–2330. [[CrossRef](#)]
30. Suganthi, S.R.L.; Basilica, M. Deep Learning Approach for Fetal health prediction. *Int. J. Sci. Res. Comput. Sci. Eng. Inf. Technol.* **2024**, *10*, 564–570. [[CrossRef](#)]
31. Salini, Y.; Mohanty, S.N.; Naga, V.; Yang, M.; Chalapathi, M.M.V. Cardiotocography data analysis for fetal health classification using machine learning models. *IEEE Access* **2024**, *12*, 26005–26022. [[CrossRef](#)]

32. Nagavibha, R.; Satheesh, M.M.; Chandy, M.P.; Prathap, B.R. An empirical and statistical analysis of fetal health classification using different machine learning algorithms. In Proceedings of the International Conference on Devices, Circuits and Systems (ICDECS), Bangalore, India, 22–23 March 2024. [\[CrossRef\]](#)
33. Mehbodniya, A.; Lazar, A.J.P.; Webber, J.; Sharma, D.K.; Jayagopalan, S.; Kousalya, K.; Singh, P.; Rajan, R.; Pandya, S.; Sengan, S. Fetal health classification from cardiotocographic data using machine learning. *Expert Syst.* **2021**, *39*, e12899. [\[CrossRef\]](#)
34. Fetal Health Classification. Kaggle. 2020. Available online: <https://www.kaggle.com/datasets/andrewmvd/fetal-health-classification> (accessed on 27 July 2026).
35. Ayres-de-Campos, D.; Bernardes, J.; Garrido, A.; Marques-de-Sá, J.; Pereira-Leite, L. SisPorto 2.0: A program for automated analysis of cardiotocograms. *J. Matern. Fetal Med.* **2000**, *9*, 311–318. [\[CrossRef\]](#)

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.