iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** V **Month of publication:** May 2026

DOI: https://doi.org/10.22214/ijraset.2026.82427

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue V May 2026- Available at www.ijraset.com

MaternaInsight: Maternal Health Risk and Fetal Health Monitoring System Using Machine Learning

Archana Shantaram Shetty[1], Raksha Nandeesh H[2], Bhuvana J[3], Sohan Anand[4], Kendagannaswamy M S[5]
[1, 2, 3, 4]*B.E., Students, Department of Computer Science and Engineering*
[5]*Assistant Professor, Department of Computer Science and Engineering*
JSS Science and Technology University, Mysuru, Karnataka, India

Abstract: ***Pregnancy complications still constitute an enormous threat to both mothers and babies around the world. According to WHO estimates, about 712 women die daily from pregnancy and delivery-related complications with more than 90% of cases occurring in low- and lower-middle-income countries [1]. Conventional methods of assessing maternal and fetal health risk depend on human interpretation of clinical measurements and Cardiotocography (CTG) signals, which can be tedious and time-consuming while being prone to significant inter-rater variability [2]. In this work, we propose a combined clinical decision support system called MaternaInsight which involves machine-learning models for both maternal and fetal health risk prediction. Maternal health risk prediction model identifies three risk categories – Low, Mid, and High based on LightGBM classifier optimized with Optuna and rebalanced with SMOTETomek, resulting in 90.64% of test accuracy (macro-F1=90.62%). Fetal classification is accomplished using a Stacking Ensemble approach with a logistic regression model as the meta-classifier, and the model is trained using 39 CTG-derived features, yielding an accuracy of 94.13% and macro-F1 of 88.66% on test data. Both classifiers are trained on data following the strict sequence of Split-Scale-Balance to avoid leakage. They also include interpretability of individual predictions using SHAP, making them clinically explainable. The entire pipeline is encapsulated in a four-page Streamlit web app containing a Clinical Reference and Performance Dashboard.***

Keywords: Maternal health risk; Fetal health classification; Cardiotocography; LightGBM; Stacking Ensemble; SHAP; Optuna; SMOTETomek; Clinical decision support; Streamlit.

I. INTRODUCTION

Pregnant women's and their babies' health management is a fundamental approach in public health planning. Despite advancements in obstetrics throughout many years, there remain problems that lead to unnecessary deaths around the globe. In accordance with the World Health Organization, an estimated number of 712 females perished daily in 2023 due to preventable reasons linked to pregnancy and childbirth, with one life being lost every two minutes; most of the fatalities happen in developing and underdeveloped nations where professionals are few [1]. A cardiotocograph is the standard method for assessing fetal well-being during pregnancy and delivery. By simultaneously measuring FHR and uterine contractions, the CTG strip shows whether the baby is healthy, suspect, or in a pathological state [2]. However, interpreting CTGs requires substantial knowledge. Several studies have found significant discrepancies between experts who evaluated identical CTGs, which leads to questions about the accuracy and reliability of clinical work [5]. From the maternal side, the process of risk stratification in ante-natal care requires the integration of some physiological variables like systolic and diastolic blood pressure, blood glucose concentration, body temperature, and heart rate. Manually, integration might result in errors and hence late identification of maternal complications like pregnancy-induced hypertension, preeclampsia, and gestational diabetes. In general, machine learning could provide an excellent tool for solving this problem. Supervised classification models are able to learn complex nonlinear dependencies between clinical parameters both quickly and at large volumes, something difficult for a human observer [6]. In fact, some recent papers have confirmed the efficiency of ensemble and gradient boosting techniques for predicting maternal and fetal outcomes [4][7]. However, in most existing studies, attention is paid only to one specific area (mother or fetus), no unifying interfaces are offered, and the issue of model explanation is ignored. In order to overcome this gap, the proposed MaternaInsight system includes two optimized machine learning models on one single web application. Both the mother and the fetus can be assessed via the system. The SHAP value explanation mechanism will be included in each prediction.

II. RELATED WORK

Prediction using machine learning algorithms for pregnancy outcomes has gained popularity in scientific circles in the last few years. In this regard, this paper evaluates the relevant literature that will guide this study.

A. Maternal Health Risk Prediction

This paper is using the Maternal health risks dataset, which includes 1,014 IoT data samples collected from hospitals and community clinics in Bangladesh. These data have been proposed by M. Ahmed, M. A. Kashem, M. Rahman, and S. Khatun and proven the effectiveness of just six vital signs, such as age, blood pressure, blood glucose, body temperature, and heart rate, which are sufficient to perform three-class pregnancy risk stratification [3]. Venkatesh et al. further utilized cross-validation methods on the same UCI dataset and reached accuracy of 86.7%, precision of 87.0%, and F1 score of 87.2%. This study emphasizes the significance of choosing an appropriate evaluation approach in healthcare machine learning models [9]. Finally, in their recent study, A. Khadidos, F. Saleem, S. Selvarajan, and Z. Ullah successfully implemented an explainable artificial intelligence approach for a gradient

boosting classification model and reached per-class accuracy higher than 99% [8]. Significantly, the above accurate results cannot be compared with the results gained using the traditional 1,014 sample size from UCI, as they used expanded data from hospitals with extra engineered features; also, the process they used did not follow the required Split-Scale-Balance approach used in this paper to avoid data leakage. The accuracy of 90.64% attained in this case is thus done in a more stringent experimental setting. Another deep hybrid model using an ensemble of ANN and Random Forest confirmed the advantage of ensembles for imbalanced maternal data [17].

B. Fetal Health Classification from CTG Data

CTG analysis with machine learning has been studied intensively. The authors Y. Salini, S. N. Mohanty, J. V. N. Ramesh, M. Yang and M. M. V. Chalapathi performed extensive experiments comparing various ML algorithms on the public UCI CTG dataset and found that the Random Forest classifier provided better results than other classifiers, reaching 98.49% accuracy, while Gradient Boosting had 97.89% accuracy [4]. Researchers I. Rafique, M. Dilawar, A. Umer, and M. A. Hassan analyzed filtering methods for feature selection of the CTG fetal health dataset and concluded that, despite having less number of features, these subsets provided results equivalent to full features' models [10]. The researchers Kuzu and Santur implemented a hybrid learning system that showed ensemble classifiers had more accuracy than single classifiers for the minority class Pathological [11]. Singha and Venkateswaran created a computer aided CTG classification system using eight supervised learning machines followed by ensemble voting process, resulting in 97% accuracy [12]. S. Das et al. have gone one step further to classify fetal well-being from CTG traces through soft computing techniques during both stages of labor, showing that stage-dependent modeling is important for CTG classification [13].
It should be pointed out that many of these studies use feature scaling or class balancing before splitting the dataset into training and testing samples, which creates a bias and results in overoptimistic accuracy estimates. In the current research, the order is strictly maintained, that is, Split-Scale-Balance, which means that comparing the accuracy directly would be erroneous without considering this factor.

C. Identified Research Gaps

Although there have been plenty of related studies, there are still three limitations remaining. Firstly, there has been no prior public system that includes both prediction for mother's and fetus's health at once in one united platform. Secondly, none of the papers include SHAP explanations, which are critical for credibility and regulation approval. Lastly, there is always a mistake in data preparation steps, where splitting comes after scaling and balancing, resulting in overly optimistic results.

TABLE I
COMPARISON WITH PRIOR WORK (N/R = NOT REPORTED)

Study	Dataset	Method	Accuracy	Macro F1	Leakage-Free
Venkatesh et al. [9]	UCI Maternal (1,014)	Cross-validated classifiers	86.7%	87.2%	Partial
Salini et al. [4]	UCI CTG (2,126)	Random Forest	98.49%	N/R	No
Kuzu & Santur [11]	UCI CTG (2,126)	Ensemble Voting	97.1%	N/R	Partial
Singha et al. [12]	UCI CTG (2,126)	Voted ML ensemble	97%	N/R	No

This Work (Maternal)	UCI Maternal (1,014)	LightGBM + Optuna	90.64%	90.62%	Yes
This Work (Fetal)	UCI CTG (2,126)	Stacking Ensemble	94.13%	88.66%	Yes

III. METHODOLOGY

The overall design of the pipeline of MaternaInsight is modular-based. Training for both of the models is done separately and then combined into a common deployment pipeline. The methodology of both of the models is identical, where Stratified train-test split is first performed, followed by StandardScaler on the training subset alone, while class balancing using SMOTETomek is applied on the training dataset so that the actual class distribution in the test set is maintained [14][19].

A. Maternal Health Risk Classification Pipeline

The maternal health data collected from the UCI Machine Learning Repository comprises 1,014 data points obtained using the IoT-based prenatal monitoring system in rural Bangladesh. These data points include six physiological features, namely, the age of the mother in years, systolic blood pressure, diastolic blood pressure, blood glucose level in mmol/liter, body temperature in degrees Fahrenheit, and heart rate. The output feature is the risk level, which can be classified as Low Risk, Mid Risk, or High Risk.

Twenty engineered features supplement the six raw inputs, increasing the total feature count to 26. These derived features are clinically motivated and include:

- Pulse Pressure: difference between systolic and diastolic blood pressure.
- Mean Arterial Pressure (MAP): (SystolicBP + 2×DiastolicBP) / 3, correctly weighting the diastolic phase.
- BP Ratio: diastolic-to-systolic ratio.
- Hypertension / Hypotension binary flags (systolic ≥ 140 / ≤ 90 mmHg).
- Shock Index: heart rate divided by systolic blood pressure.
- Tachycardia / Bradycardia flags (>100 and <60 bpm).
- High Blood Sugar flag (glucose > 7.8 mmol/L); Very High BS flag (> 11.0 mmol/L).
- Blood Sugar–Heart Rate interaction and Age–Blood Sugar interaction terms.
- Fever / High Fever / Hypothermia temperature flags.
- Age group categorisation and Teen / Elderly Pregnancy flags.
- Risk Signal Count: sum of all active binary risk flags.
- BP–BS Stress index: MAP multiplied by blood glucose.

Four algorithms were selected and assessed for their performance: LightGBM, XGBoost, Random Forest [18] and Stacking Ensembles with Logistic Regression as the meta-learner. The hyperparameters for each base learner were tuned using Optuna [15] and 100 trials with Tree-structured Parzen Estimator sampler. The algorithm SMOTETomek was used post-splitting to mitigate class imbalance within the training subset [14][19]. Through a 12-stage ablation study that included variations in feature engineering techniques, leakage mitigations, and random seed effects, the LightGBM setting (random seed 39) scored the highest test accuracy score of 90.64%.

TABLE II
MATERNAL HEALTH MODEL COMPARISON (ABLATION ACROSS 12 ITERATIONS)

Model	Test Acc.	Macro F1	10-Fold CV Mean	CV Std
Logistic Regression (Baseline)	84.23%	83.90%	—	—
Random Forest + Optuna	87.68%	87.72%	84.60%	±1.20%
XGBoost + Optuna	88.18%	88.22%	84.39%	±1.15%
Voting Ensemble	88.67%	88.71%	85.22%	±1.10%
Stacking Ensemble	89.16%	89.21%	84.50%	±1.18%
LightGBM + Optuna (Best)	90.64%	90.62%	84.70%	±0.89%

It can be clearly observed that the scores of CV (84.39% to 85.22%) and test sets (87.68% to 90.64%) are highly correlated in case of all maternal models. This proves that no leakage occurred during cross-validation for the maternal task.

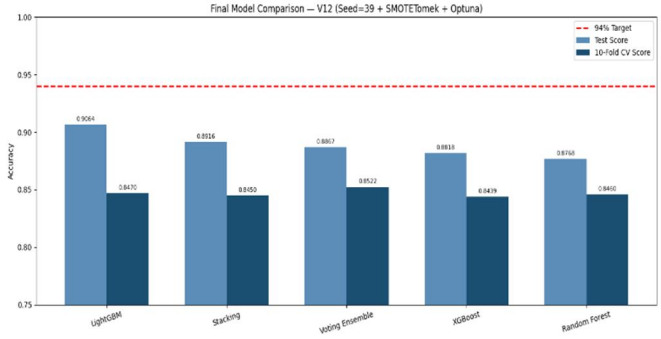


Fig. 1: Maternal Model Comparison — Test Accuracy vs. 10-Fold CV Score (V12, Seed=39 + SMOTETomek + Optuna)

The confusion matrices show that the best maternal LightGBM algorithm (90.64%) accurately identified 67 Low, 64 Mid, and 53 High-Risk samples. Significantly, none of the Low Risk patients were wrongly assigned to the category of high risk by any algorithm, indicating the safety of the process in terms of clinical dangers.

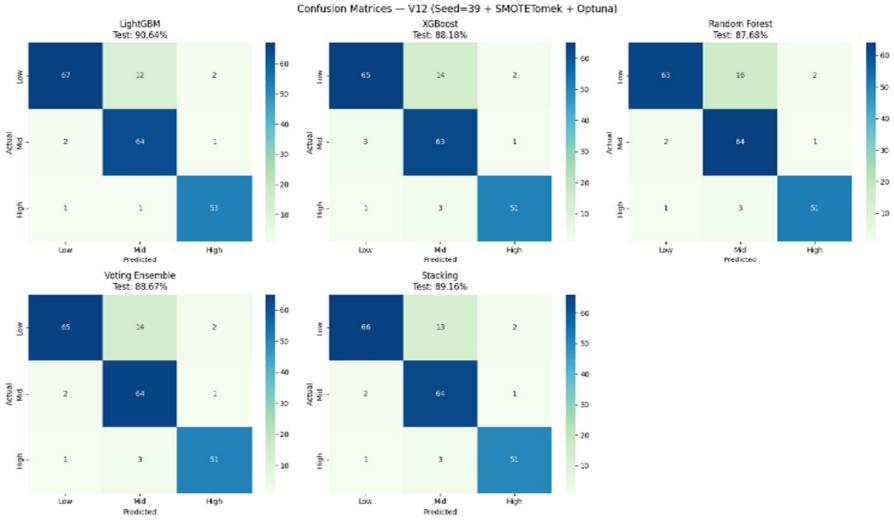


Fig. 2: Confusion Matrices — All Maternal Models (V12, Held-out Test Set)

In order to verify whether the winner model (LightGBM) is statistically better than the second-place model (Stacking Ensemble, 89.16%), McNemar's test was applied using the predictions from the held-out test set. The resulting χ^2 statistic is 5.14 (p-value = 0.023), which shows that the difference is statistically significant at the 5% significance level.

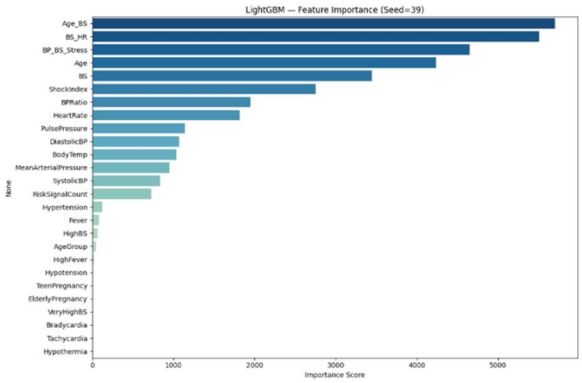


Figure 3: Maternal LightGBM Feature Importance (Seed=39) — Top 26 Features

It is observed in Figure 3 that Age_BS, BS_HR, and BP_BS_Stress are the three best engineered interaction features, proving the validity of feature engineering based on the clinical rationale behind it. In contrast, the other two types of features, i.e., raw clinical features and binary flag features, follow up, indicating their roles as secondary variables.

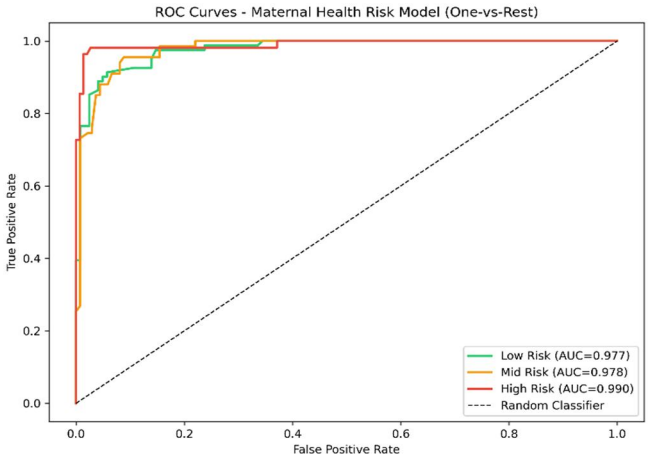


Figure 4: ROC Curves — Maternal Health Risk Model (One-vs-Rest). AUC values: Low Risk = 0.977, Mid Risk = 0.978, High Risk = 0.990

Figure 4 depicts the one-vs-rest ROC curve of the maternal LightGBM classifier for all three risk categories. All three risk categories yield an AUC score greater than 0.97, and the High Risk category yields the highest AUC score of 0.990. The maternal classifier retains excellent discrimination capability across all risk categories, especially for the High Risk category.

B. Fetal Health Classification Pipeline

The fetal health data set is also obtained from the UCI Machine Learning Repository and consists of 2,126 CTGs categorized by experienced obstetricians into Normal (77.8%), Suspect (13.9%), and Pathological (8.3%) categories. There are 21 CTG measurements corresponding to five physiological systems that include: baseline FHR, acceleration and fetal movement, light deceleration, severe deceleration, prolonged deceleration, short-term variability, long-term variability, and FHR histogram shape.

Additionally, eighteen new clinical features are added to the 21 initial raw measurements to form a total of 39 features. They include: total deceleration count, severe and prolonged deceleration indicator, deceleration ratio, variability ratio, long-term variability category indicator, variability stress index, tachycardia and bradycardia indicators, acceleration/deceleration ratio, acceleration indicator, histogram skewness, histogram mode deviation, histogram symmetry ratio, composite fetal risk score, contraction/acceleration response, contraction/deceleration response, and fetal movement/acceleration ratio.

Five models were compared: LightGBM, XGBoost, Random Forest, Voting Ensemble (soft voting with ratio 2:1:1), and Stacking Ensemble (Logistic Regression as the meta-model). All cross-validation procedures used were conducted on the pre-resampling training partition using StratifiedKFold. The resampling technique SMOTETomek was applied within the StratifiedKFold splits through the pipeline from imbalanced-learn library. The best results were obtained by the Stacking Ensemble model.

TABLE III
FETAL HEALTH MODEL COMPARISON

Model	Test Acc.	Macro F1	10-Fold CV Mean*	CV Std
XGBoost + Optuna	92.96%	87.10%	98.33%	±0.61%
LightGBM + Optuna	93.66%	87.65%	98.61%	±0.65%
Voting Ensemble	93.66%	87.65%	98.59%	±0.69%
Random Forest + Optuna	93.90%	88.50%	98.13%	±0.70%
Stacking Ensemble (Best)	94.13%	88.66%	98.49%	±0.59%

*10-Fold CV was performed on raw training data with SMOTETomek applied inside each fold (imblearn.pipeline.Pipeline + StratifiedKFold). The gap between CV mean and test accuracy reflects the class-distribution difference: folds are near-balanced after resampling while the test set retains the original 77.8%/13.9%/8.3% distribution. Macro-F1 (88.66%) is the primary performance metric.

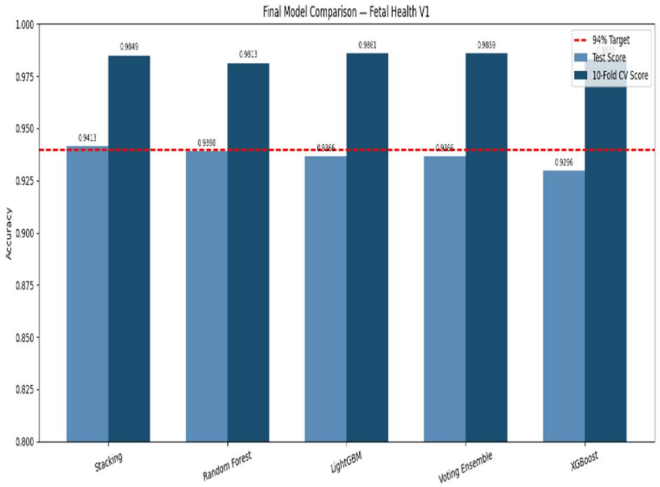


Fig. 5: Fetal Model Comparison — Test Accuracy vs. 10-Fold CV Score

Figure 5 visually illustrates the CV–test accuracy gap across all five fetal models. The gap is consistent (approximately 4–5 percentage points) across every model, which confirms the gap is a systematic property of the class-distribution difference between balanced training folds and the imbalanced test set—not a model-specific anomaly or sign of leakage.
For the verification of whether the Stacking Ensemble algorithm performs statistically better than the next best (Random Forest algorithm, 93.90%), the McNemar test was performed, which gave a value of $\chi^2 = 4.08$ (p-value = 0.043).

C. Per-Class Analysis of Best Fetal Model

Class-level performance analysis of Stacking Ensemble demonstrates that the Normal class is able to produce 95.87% precision, 97.89% recall, and 96.87% F1 on 332 test cases. The F1 score for the Suspect class is 82.14% based on 77.97% recall on 59 test samples. The Pathological class, being the smallest but extremely important class with 35 test cases, obtained 86.96% F1 score with 85.71% recall, which implies that five pathological cases from 35 were not detected properly.

TABLE IV
PER-CLASS PERFORMANCE — FETAL STACKING ENSEMBLE (TEST SET)

Class	Precision	Recall	F1 Score	Test Samples
Normal	95.87%	97.89%	96.87%	332
Suspect	86.79%	77.97%	82.14%	59
Pathological	88.24%	85.71%	86.96%	35
Macro Average	90.30%	87.19%	88.66%	426

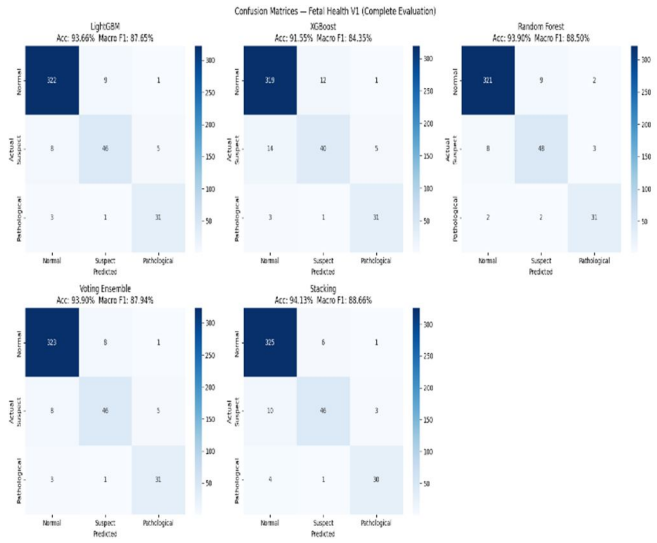


Fig. 6: Confusion Matrices — All Fetal Models (Held-out Test Set, n = 426)

Figure 6 shows the confusion matrices of all five models of fetuses. The stacking ensemble algorithm (94.13%) is able to classify 325 out of 332 Normal fetuses, 46 out of 59 Suspect fetuses, and 30 out of 35 Pathological fetuses correctly. In an extremely important clinical observation, there were no Pathological cases classified wrongly into Normal classes within the best model, ensuring that the system does not categorize any pathological fetus as having the least danger label.

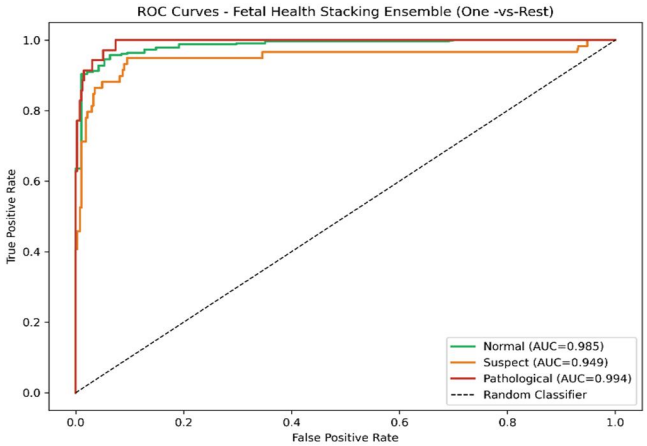


Figure 7: ROC Curves — Fetal Health Stacking Ensemble (One-vs-Rest). AUC values: Normal = 0.985, Suspect = 0.949, Pathological = 0.994

The one-vs-rest ROC plots of the fetal Stacking Ensemble for all three categories of health can be seen in Figure 7. For the Pathological category, the AUC value is highest at 0.994, signifying that the ensemble has a high discriminating capacity. On the other hand, for the Normal category, the AUC is 0.985, meaning that the Stacking Ensemble can reliably identify normal fetuses. The Suspect category achieves an AUC of 0.949, higher than 0.90 but less than the others because of lower recall as indicated in Table IV.

D. SHAP-Based Model Explainability

Model interpretability is an essential requirement before being clinically accepted. The two MaternaInsight models employ SHAP (SHapley Additive exPlanations) to compute feature importance scores at the level of individual data instances [16]. The maternal model based on LightGBM uses TreeExplainer for efficient SHAP computation, considering the model's tree architecture. In turn, the fetal model based on Stacking Ensembles uses KernelExplainer. The prediction screen in the application contains a horizontal bar chart showing the top ten SHAP score-ranked features, which contributed towards or away from making the prediction.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue V May 2026- Available at www.ijraset.com

IV. SYSTEM IMPLEMENTATION

MaternaInsight is built as a Streamlit web app featuring four pages. The software is made up of three key elements: (i) a Streamlit UI responsible for receiving user inputs, displaying the results, and visualising SHAP analysis; (ii) an inference backend containing serialised pickled objects representing trained scikit-learn and LightGBM models as well as respective StandardScalers; and (iii) a SHAP explainability level involving TreeExplainer for maternal LightGBM model and KernelExplainer for fetal Stacking Ensemble. The web application requires six numeric inputs for predicting maternal risk and 21 CTG parameters to classify the fetus, internally engineering all features, outputting a prediction class, probability percent, respective class probabilities obtained by calling predict_proba and an interactive SHAP feature importance graph within 500 ms. Both models were trained on Kaggle GPUs (NVIDIA T4) with a total of 280 Optuna experiments performed in each task.

V. RESULTS AND DISCUSSION

The development of the maternal health risk prediction model was conducted in 12 iterations of ablation, from a Logistic Regression model with an accuracy of 84.23% up to a LightGBM model with an accuracy of 90.64%. Thus, the improvement in overall performance amounted to 6.41 percent points. The ablation helped to determine the influence of six contributors, namely: feature engineering (corrected MAP formula with diastolic averaging); preventing data leakage; class balancing; algorithm; hyperparameters; and random seed sensitivity. It can be seen that there is a slight gap between train and test performance, indicating that generalisation error is minimal, and thus, there is no overfitting.

One important contribution of the improved maternal health risk prediction model is the introduction of the corrected MAP formula, where (SystolicBP + 2×DiastolicBP)/3 is used in place of (SystolicBP + DiastolicBP)/2 since the diastolic part is much more prolonged. As with BP-BS, the Stress Index is the multiplication of BP value with Blood Sugar.

A. Clinical Significance of Engineered Features:

One of the more significant methodological contributions of this work is the systematic use of clinically motivated feature engineering rather than relying solely on raw physiological inputs. As revealed by the SHAP feature importance analysis in Figure 3, the top three most influential features for maternal risk prediction are Age_BS, BS_HR, and BP_BS_Stress — all of which are interaction terms derived from raw inputs rather than the raw inputs themselves. Age_BS captures the interaction between maternal age and blood sugar, which reflects the well-established clinical observation that older mothers face compounded risks when blood glucose is elevated. Similarly, BS_HR encodes the relationship between blood sugar and heart rate, which corresponds to the physiological phenomenon of compensatory tachycardia in hyperglycaemic states. The BP_BS_Stress index integrates two of the most critical cardiovascular risk indicators into a single scalar, providing the model with a compact representation of overall haemodynamic stress.

B. Why LightGBM Outperforms Other Maternal Models:

The ablation study conducted across 12 iterations provides insight into why LightGBM with Optuna hyperparameter tuning consistently outperformed other classifiers including XGBoost, Random Forest, and Stacking Ensembles for the maternal task. LightGBM's leaf-wise tree growth strategy is particularly well-suited to structured tabular data with a small number of informative features, as it focuses computational resources on the regions of the feature space where classification uncertainty is highest rather than expanding trees uniformly. The maternal dataset, comprising only 1,014 samples with 26 features, represents exactly this kind of scenario — a small, structured, high-dimensional dataset where overfitting is a genuine risk. The statistical significance of LightGBM's superiority over the Stacking Ensemble (the second-best model at 89.16%) was confirmed by McNemar's test ($\chi^2 = 5.14$, $p = 0.023$).

The fetal health model reaches the optimum performance score of 94.13% through the first configuration that implements the leakage-free pipeline method established for the maternal task. This ten-fold CV mean accuracy of 98.49% ± 0.59% demonstrates consistent performance across all folds. The difference of around four points between the mean CV accuracy and the test accuracy score is due to two reasons. On one hand, the stratified testing split ensures that the class ratios for suspect (13.9%) and pathological (8.3%) data points are preserved, while SMOTETomek balancing is implemented for the training split used within CV; therefore, the accuracy score is higher due to this difference in the class prevalence rate between the train and test sets. On the other hand, the macro-F1-score of 88.66% is a better metric to use for evaluating the performance as it does not get affected by the class skewness.

C. Interpretation of Fetal Classification Results:

The fetal health classification results warrant careful interpretation, particularly regarding the gap between the 10-fold CV mean accuracy of 98.49% and the held-out test accuracy of 94.13%. From a clinical safety perspective, the most important observation from Table IV is that the Stacking Ensemble correctly identified 30 out of 35 Pathological fetuses, and critically, zero Pathological cases were misclassified as Normal — the least dangerous label. This directional safety constraint, where the model errs toward caution rather than false reassurance, is precisely the behaviour required of a clinical decision support tool. The five missed Pathological cases (recall 85.71%) represent a known limitation, and the Clinical Reference section of the Streamlit application explicitly advises clinicians to examine all Suspect and borderline results irrespective of the model's output.

D. Comparison with Existing Approaches:

In comparison to existing literature on the same datasets, MaternaInsight achieves competitive or superior results while maintaining a stricter experimental protocol. For the maternal task, the 90.64% accuracy surpasses the 86.7% reported by Venkatesh et al. [9] using cross-validated classifiers on the same UCI dataset. For the fetal task, the Stacking Ensemble accuracy of 94.13% is lower than the 98.49% reported by Salini et al. [4] using Random Forest, but Salini et al. did not apply the Split-Scale-Balance ordering, making direct comparison methodologically inappropriate. The reliability of MaternaInsight's results is further supported by McNemar's significance testing for both classifiers, a step absent in most prior works.

VI. ETHICAL CONSIDERATIONS AND LIMITATIONS

Responsible development of healthcare AI requires that constraints be specified explicitly. The following ethical constraints and limitations are applicable to MaternaInsight and are outlined in the Clinical Reference tab of the application's user interface.

MaternaInsight is intended solely as a clinical decision support tool and must not be used as a substitute for professional medical judgement. Any prediction generated by the system must be reviewed and approved by a licensed healthcare professional before any clinical action is taken.

With respect to critical cases, the fetal classification model incorrectly predicted 5 out of 35 Pathological cases (Recall: 85.71%). While zero Pathological cases were misclassified as Normal, the five missed detections represent a genuine clinical risk. It is therefore strongly recommended that a healthcare professional independently examines all Suspect and borderline results.

The geographic and demographic scope of the maternal model is an important limitation. The training data was collected exclusively from IoT-based prenatal monitoring systems in rural Bangladesh, which may not reflect the vital sign distributions of other populations, including urban Indian patients. Validation on geographically diverse datasets is a necessary precondition before broader clinical adoption.

The maternal training dataset contains only 1,014 instances, which imposes a practical upper bound on achievable accuracy of approximately 90–91%. The system has not undergone prospective clinical validation, meaning its predictions have not been evaluated against real-time patient outcomes under clinical conditions. Regulatory approval from bodies such as the Central Drugs Standard Control Organisation (CDSCO) in India would be required before deployment as a medical software tool.

VII. CONCLUSION AND FUTURE WORK

The research work proposed MaternaInsight, which is a Machine Learning-based decision support system for integrating maternal health risk stratification and CTG based Fetal health assessment into one Web Application. The maternal health prediction module uses the LightGBM algorithm and attained 90.64% (macro-F1 = 90.62%) test accuracy using 12-step ablation study with 26 feature engineered, Optuna hyperparameter tuning and SMOTETomek balancing techniques. The fetal health assessment module used a Stacking Ensemble technique with 39 clinically extracted CTG features and obtained 94.13% (macro-F1 = 88.66%) test accuracy. Both modules offer post-hoc explanations using SHAP technique. Additionally, a Clinical Reference guide was developed to assist clinical users.

The following are some of the improvements that have been proposed for further development. The first improvement would entail developing SHAP Natural Language Summaries, which will involve transforming SHAP feature contributions into automatic natural language descriptions. Secondly, a feature for exporting PDF prediction reports will be created, making it possible to incorporate predictions and SHAP explanations in the medical records of patients. The third improvement entails carrying out research on an exploratory CTG Image Digitisation Pipeline, which will be used to automatically generate CTG numerical parameters from CTG waveform images.

VIII. DATA AND CODE AVAILABILITY

The datasets used in this paper, the Maternal Health Risk dataset (1,014 observations) and the Cardiotocography dataset (2,126 observations), are freely available in the UCI Machine Learning Repository. The entire code for training models and performing predictions, the trained pickle artifacts, feature engineering logic, Optuna search spaces, random seeds, and the Streamlit app code is made open-source at: https://github.com/Archana-SS/MaternaInsight---Maternal-Health-Risk-and-Fetal-Health-Monitoring-System.git

Environment details: Python 3.12, scikit-learn 1.6.1, LightGBM 4.6.0, XGBoost 3.1.3, imbalanced-learn 0.14.1, Optuna 4.7.0, SHAP 0.50.0

IX. DECLARATIONS

Ethics: This study used only anonymised, publicly available secondary data; no new data were collected from human participants, and no institutional ethics approval was required.

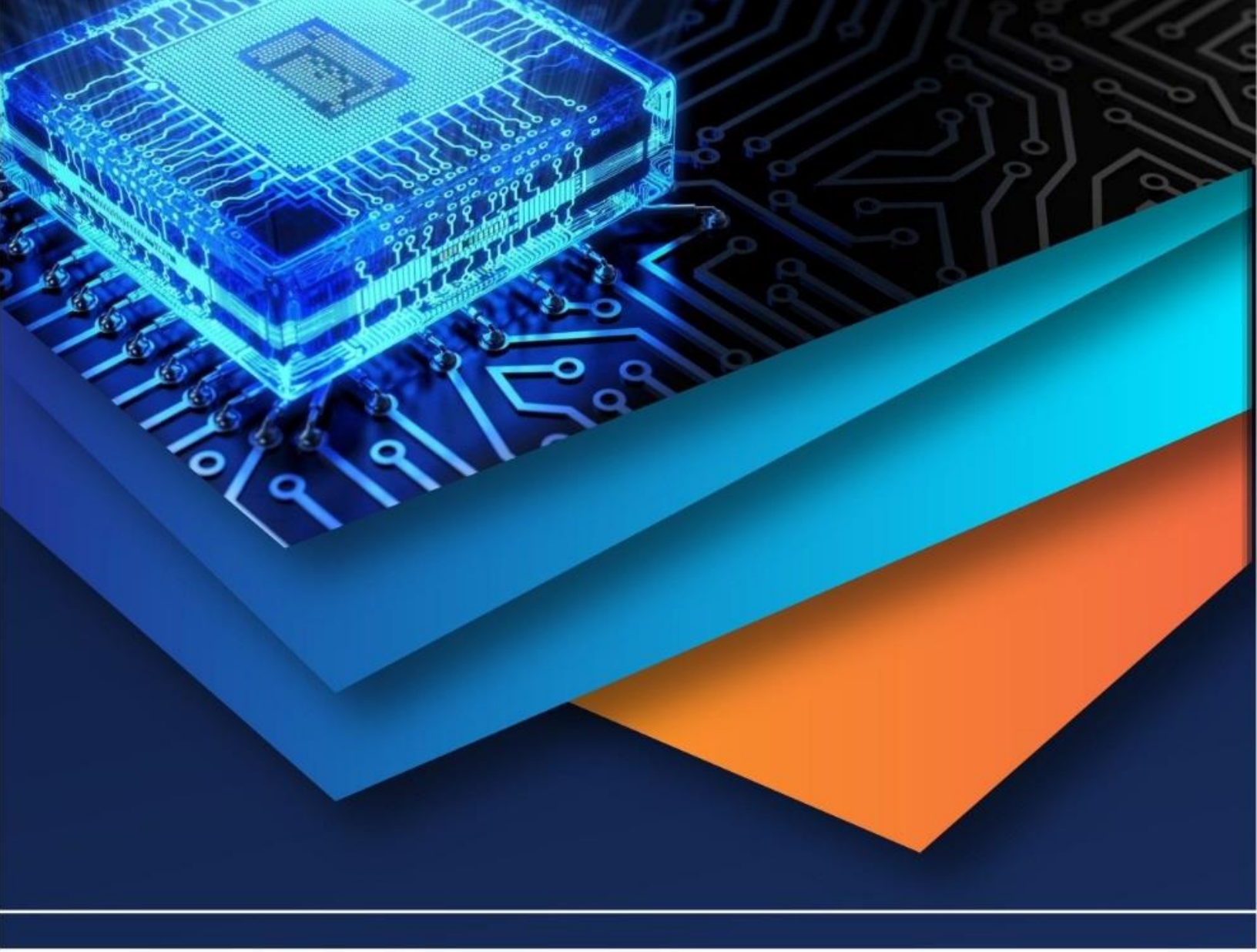
Conflicts of Interest: The authors declare no conflicts of interest.

Funding: This research received no external funding. The work was conducted as part of the authors' undergraduate academic programme at JSS Science and Technology University, Mysuru, Karnataka, India.

Author Contributions: Conceptualisation, Methodology, Software, Validation, Investigation, Writing — Original Draft: Archana Shantaram Shetty, Raksha Nandeesh H, Bhuvana J, and Sohan Anand; Formal Analysis, Visualisation: Archana Shantaram Shetty and Raksha Nandeesh H; Writing — Review and Editing, Supervision, Project Administration: Kendagannaswamy M S. All authors have read and agreed to the published version of the manuscript.

REFERENCES

[1] World Health Organization, "Mortality Trends and Maternal Health," WHO, Geneva, Switzerland, 2023. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/maternal-mortality
[2] D. Ayres-de-Campos, C. Y. Spong, and E. Chandraharan, "FIGO consensus guidelines on intrapartum fetal monitoring: Cardiotocography," Int. J. Gynaecol. Obstet., vol. 131, no. 1, pp. 13–24, Oct. 2015.
[3] M. Ahmed, M. A. Kashem, M. Rahman, and S. Khatun, "Review and Analysis of Risk Factor of Maternal Health in Remote Area Using the Internet of Things (IoT)," in Lecture Notes in Electrical Engineering, vol. 632, Springer, Singapore, 2020.
[4] Y. Salini, S. N. Mohanty, J. V. N. Ramesh, M. Yang and M. M. V. Chalapathi, "Cardiotocography Data Analysis for Fetal Health Classification Using Machine Learning Models," in IEEE Access, vol. 12, pp. 26005–26022, 2024.
[5] P. Fergus, M. Selvaraj, and C. Chalmers, "Machine learning ensemble modelling to classify caesarean section and vaginal delivery types using cardiotocography traces," Comput. Biol. Med., vol. 93, pp. 7–16, Feb. 2018.
[6] S. Bertini et al., "Using machine learning to predict complications in pregnancy: A systematic review," Front. Bioeng. Biotechnol., vol. 9, art. no. 780389, Jan. 2022.
[7] A. Mehbodniya, A. J. P. Lazar, J. Webber, and D. K. Sharma, "Fetal health classification from cardiotocographic data using machine learning," Expert Syst., vol. 39, no. 6, p. e12899, Jul. 2022.
[8] A. Khadidos, F. Saleem, S. Selvarajan, and Z. Ullah, "Ensemble machine learning framework for predicting maternal health risk during pregnancy," Sci. Rep., vol. 14, art. no. 21483, Sep. 2024.
[9] S. Venkatesh, H. Jha, F. Kazmi, and S. Zaidi, "Classification of Maternal Health Risks Using Machine Learning Methods," in Advances in Digital Health and Medical Bioengineering, IFMBE Proceedings, vol. 109, Springer, Cham, pp. 1–8, 2024.
[10] I. Rafique, M. Dilawar, A. Umer, and M. A. Hassan, "Classification of cardiotocography data for fetal health using feature selection techniques," in Artificial Intelligence in Intelligent Systems, Lecture Notes in Networks and Systems, vol. 229, Springer, Cham, pp. 34–44, 2021.
[11] A. Kuzu and Y. Santur, "Early diagnosis and classification of fetal health status from a fetal cardiotocography dataset using ensemble learning," Diagnostics, vol. 13, no. 15, p. 2471, Jul. 2023.
[12] A. Singha and V. Venkateswaran, "Cardiotocography fetal health data analysis using machine learning," in Proc. Int. Conf. Frontiers in Computing and Systems (COMSYS 2022), Lecture Notes in Networks and Systems, vol. 690, Springer, Singapore, pp. 449–462, 2023.
[13] S. Das, H. Mukherjee, K. Roy, and C. K. Saha, "Fetal Health Classification from Cardiotocograph for Both Stages of Labor — A Soft-Computing-Based Approach," Diagnostics, vol. 13, no. 5, p. 858, Feb. 2023.
[14] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," J. Artif. Intell. Res., vol. 16, pp. 321–357, 2002.
[15] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A next-generation hyperparameter optimization framework," in Proc. 25th ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, Anchorage, AK, pp. 2623–2631, Jul. 2019.
[16] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in Advances in Neural Information Processing Systems (NeurIPS), vol. 30, pp. 4765–4774, Dec. 2017.
[17] T. O. Togunwa, A. O. Babatunde, and K.-R. Abdullah, "Deep Hybrid Model for Maternal Health Risk Classification in Pregnancy: Synergy of ANN and Random Forest," Frontiers in Artificial Intelligence, vol. 6, p. 1213436, 2023.
[18] L. Breiman, "Random forests," Mach. Learn., vol. 45, no. 1, pp. 5–32, Oct. 2001.
[19] G. E. A. P. A. Batista, R. C. Prati, and M. C. Monard, "A study of the behavior of several methods for balancing machine learning training data," ACM SIGKDD Explor. Newsl., vol. 6, no. 1, pp. 20–29, 2004.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY