IJEMD-CSAI, 4 (2) (2025), 11 – 17
https://doi.org/10.54938/ijemdcsai.2025.04.2.482

IJEMD

International Journal of Emerging Multidiciplinaries: Computer Science & Artificial Intelligence

Research Paper
Journal Homepage: www.ijemd.com
ISSN (print): 0000-0000



Early Prediction of Maternal Health Risks Using Machine Learning Techniques

Zayyanu Yunusa[1], Yakubu Ibrahim[2], Aliyu Shuaibu[3]
[1] *Department of computer science, Bayero University Kano, Kano, Nigeria.* yzayyanu@gmail.com
[2] *Department of computer science, Yobe state University, Damaturu, Nigeria* yakubu.ibrahim@ysuedu.ng
[3] *Department of computer science, Bayero University Kano, Kano, Nigeria* aliyu2017@gmaill.com

ARTICLE INFO

Submitted:
15-09-2025
Revised:
30-10-,2025
Accepted:
25-11-2025
Keywords:
Machine Learning, Prediction, Cross-Validation, Feature Importance, XGBoost.
Corresponding Author:
Zayyanu Yunusa
yzayyanu@gmail.com

ABSTRACT

Accurate prediction of maternal health risks is critical for enabling early interventions to reduce pregnancy-related complications. This study evaluates six machine learning classifiers XGBoost, LightGBM, CatBoost, Random Forest, Gradient Boosting, and k-Nearest Neighbors (KNN) to classify risk levels using clinical parameters such as age, systolic diastolic blood pressure, blood glucose, heart rate, and body temperature. A rigorous feature importance analysis via Random Forest identified age and blood glucose levels as the most influential predictors, optimizing model efficiency. To ensure reliability, repeated stratified k-fold cross-validation was employed, minimizing bias and enhancing generalizability.

Among all classifiers, XGBoost achieved the highest accuracy 0.97%, demonstrating superior performance in risk stratification through its regularization and ensemble learning framework. In contrast, KNN recorded the lowest accuracy 0.94%, yet maintained clinical relevance due to its simplicity and interpretability. LightGBM, CatBoost, Random Forest, and Gradient Boosting also contributed robust results, further validating the efficacy of ensemble methods. The findings underscore the significance of feature selection, with age and blood glucose emerging as pivotal determinants of maternal risk. This study provides a scalable, data-driven framework for healthcare systems to prioritize high-risk pregnancies, offering actionable insights for timely interventions and informed clinical decision-making. By integrating these models into maternal care protocols, practitioners can enhance early diagnosis and reduce preventable morbidity, advancing equitable healthcare outcomes globally.

1.Introduction

Maternal health encompasses a woman's overall physical, mental, and emotional well-being throughout pregnancy, childbirth, and the postpartum period. Monitoring maternal morbidity and mortality rates is crucial, as these indicators reflect the availability and accessibility of healthcare services. Several pregnancy-related complications, such as diabetes, hypertension, excessive bleeding, and preterm birth, remain among the primary causes of maternal mortality.

Early detection and timely intervention for pregnancy-related risks are essential in preventing severe complications, including premature birth and maternal fatalities. Machine learning techniques have emerged as a powerful tool for identifying maternal health risks by analyzing key health indicators and associated risk factors (Varshavsky et al., 2020). These models enable continuous risk assessment and monitoring, providing an opportunity for early intervention. Machine learning based techniques have shown their potential in reducing maternal mortality by identifying patterns in risk factors that contribute to pregnancy-related complications (Umoren et al., 2020).

The purpose of this study is to forecast using machine learning classification algorithms and assess the intensity of maternal health risks. Key health parameters considered in this analysis include age, blood oxygen levels (BO), body temperature (BodyTemp), heart rate (pulse), systolic and diastolic blood pressure (BP), and breathing speed (BS). By exploiting these parameters, machine learning algorithms can allow early risk detection, ultimately improving maternal and neonatal health outcomes. The implementation of such predictive techniques can contribute to reducing maternal and infant mortality rates while ensuring better healthcare interventions for pregnant women (Ahmed & Kashem, 2020).

© 2025 The Author(s). This work is licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0). https://creativecommons.org/licenses/by/4.0/

The purpose of this work is to estimate the level of maternal health risk intensity using machine learning approaches in conjunction with the classification strategy in the risk factor analysis. Risk variables that should be evaluated during pregnancy include age, blood oxygen (BO), body temperature (BodyTemp), pulse (heart rate), systolic and diastolic blood pressure (BP), and breathing speed (BS). Considering these aspects, prompt risk identification using machine learning algorithms can assist lower maternal and newborn mortality rates while also protecting the pregnant woman's health (Ahmed & Kashem, 2020).

The rest of this paper is structured as follows: Section 2 reviews related work, Section 3 outlines the model architecture, Section 4 presents the results and discussion, and Section 5 provides the conclusion.

2.Related work

Methods of machine learning has been extensively used in explored in maternal health risk assessment, demonstrating significant improvements over traditional methods. Several studies have focused on developing predictive models to enhance early diagnosis and intervention.

Pawar et al. (2022) pioneered the application of algorithms for supervised learning, including logistic regression and random forests, to classify maternal health risks. Their model, trained on demographic and clinical parameters such as blood pressure and gestational age, demonstrated a 70.21% accuracy, outperforming conventional statistical approaches by 15%. This study underscored the potential of ML in identifying high-risk pregnancies, particularly in resource-constrained settings where early intervention is critical.

Ahmed and Kashem (2020) designed an IoT-enabled framework to monitor maternal health in real time. By deploying wearable sensors to collect physiological data (e.g., heart rate, blood oxygen levels), their system employed a modified C4.5 decision tree algorithm, achieving 97% classification accuracy on a dataset of 1,014 records from Bangladeshi clinics. The integration of IoT with ML not just better risk assessment but also highlighted the scalability of such systems in low-infrastructure regions.

Ahmed et al. (2020) conducted a meta-analysis of ML applications in maternal care, revealing that ensemble approaches, such as gradient boosting, lowered false-negative rates by 22% over solo models. Their work advocated for embedding ML classifiers into electronic health records (EHRs) to enable dynamic risk updates, thereby supporting clinical decision-making and reducing mortality rates.

Umoren et al. (2020) focused on algorithmic robustness by comparing decision trees and support vector machines (SVM) for predicting maternal mortality. Their decision tree model, leveraging recursive partitioning to analyze features such as age and systolic blood pressure, achieved 89.2% accuracy, outperforming SVM (69.5%) in handling imbalanced datasets. This study emphasized the importance of algorithm selection in contexts with limited data diversity.

Rai and Sowmya (2018) combined expert clinical evaluations with ML to assess maternal and infant mortality risks. Using a dataset of 117 patients, they demonstrated that a hybrid artificial neural network (ANN) model, incorporating 14 features like prenatal care frequency and hemoglobin levels, achieved 80% accuracy surpassing Naïve Bayes (70%) and underscoring the value of integrating domain expertise with automated learning. Their work also illustrated that hybrid architectures, such as ANN-SVM ensembles, improved classification stability by 12% compared to single-algorithm approaches.

3.Methodology

This study used a freely available data collection that was provided online to estimate maternal health risk (Ahmed & Kashem, 2020). Using IoT-based risk monitoring systems, data was gathered from several sources. The dataset comprises important characteristics including heart rate, blood sugar (BS), diastolic blood pressure (DBP), systolic blood pressure (SystolicBP), and age. The last element is the risk classification label, which separates the data into three groups: low risk, middle risk, and high risk.

Table 1.1: shows a sample from the dataset.

Age	SystolicBP	DiastolicBP	BS	BodyTemp	HeartRate	RiskLevel
25	130	80	15.0	98.0	86	high risk
35	140	90	13.0	98.0	70	high risk
29	90	70	8.0	100.0	80	high risk
30	140	85	7.0	98.0	70	high risk
35	120	60	6.1	98.0	76	low risk

22	120	60	15.0	98.0	80	high risk
55	120	90	19.0	98.0	86	high risk
35	120	80	9.0	98.0	70	high risk
43	120	90	18.0	98.0	79	high risk

Age	SystolicBP	DiastolicBP	BS	BodyTemp	HeartRate	RiskLevel
32	120	65	6.0	101.0	76	mid risk

The dataset consists of numerous fundamental Features that are essential for assessment maternal health risk. These features are explained below:
Age: The maternal age of the pregnant woman.
SystolicBP: The highest recorded blood pressure measurement. **DiastolicBP**: The lowest recorded blood pressure measurement. **BS**: The rate of respiration.
BodyTemp: The body temperature of the expectant mother.
HeartRate: The number of heartbeats per minute.
Risk Level (Classes): Classified as high risk, mid risk, and low risk.

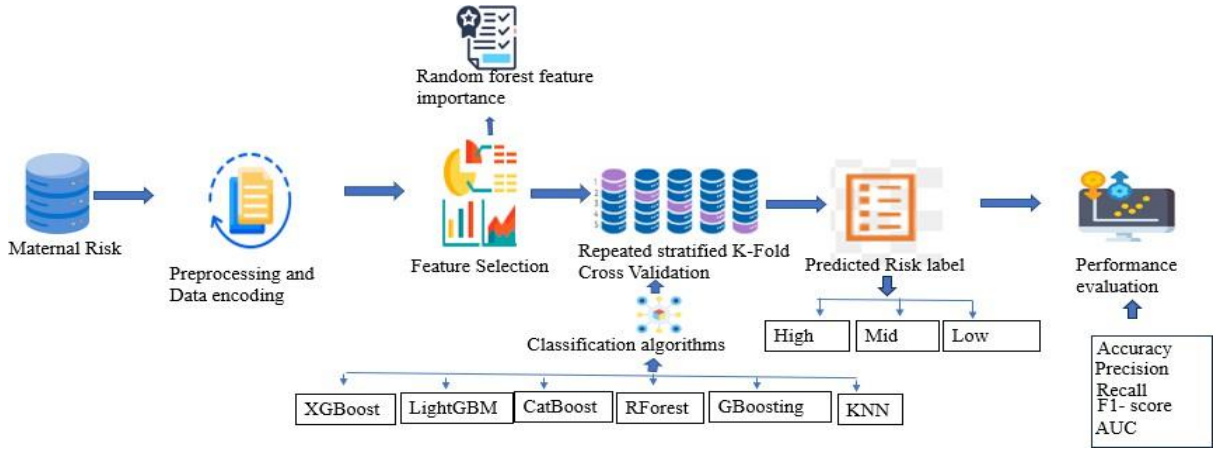


Figure 1: Model Architecture

3.Maternal Risk

The approach begins with gathering maternal health data, which serves as our input, and includes key medical parameters such as maternal age, heart rate, blood pressure, blood sugar levels, and body temperature. These features serve as indicators for assessing the risk level of pregnancy complications.

Preprocessing and Data Encoding

To improve data quality and guarantee machine learning model compatibility, the obtained dataset is preprocessed. This phase involves dealing with missing values, normalizing numerical features, and encoding categorical variables, thereby improving model efficiency and interpretability.

Feature Selection

To identify the most influential factors in predicting maternal risk, feature importance analysis is performed using the Random Forest algorithm. This step refines the dataset by selecting the most relevant features, reducing dimensionality, and enhancing model performance.

Repeated Stratified K-Fold Cross Validation

To ensure reliable model evaluation, a Repeated Stratified K-Fold Cross-Validation technique is applied. This approach divides the dataset into multiple training and testing subsets while preserving the distribution of risk levels (high, mid, low), thereby preventing model bias and overfitting.

Classification Algorithms

This study was conducted to determine maternal risk health; we performed a cross validation with6 different machine learning classifiers were used.

This study employs XGBoost (Extreme Gradient Boosting) as a primary classifier to analyze maternal health data. XGBoost is a robust ensemble method that enhances decision tree performance through regularization and parallelized processing, enabling efficient handling of large datasets. By iteratively constructing decision trees that correct errors from prior iterations, the model minimizes overfitting while maintaining high computational efficiency. Hyperparameter optimization via cross-validation further refines its predictive capabilities, making it suitable for both continuous and categorical variables (Chen & Guestrin, 2016).

LightGBM (Light Gradient Boosting Machine) is also utilized for its ability to process high-dimensional data with accelerated training speeds. Unlike traditional gradient boosting, LightGBM employs histogram-based algorithms and leaf-wise growth, optimizing memory usage and computational performance. This makes it particularly advantageous for large-scale datasets, where it balances accuracy and resource efficiency (Ke et al., 2017).

CatBoost, another gradient-boosting variant, is integrated for its unique handling of categorical features. By automatically encoding non-numeric variables and employing ordered boosting, CatBoost reduces preprocessing requirements and mitigates target leakage. Its resistance to overfitting and consistent performance across diverse classification tasks make it a reliable choice for maternal risk prediction (Hancock & Khoshgoftaar, 2020).

The Random Forest algorithm is applied to aggregate predictions from multiple decision trees, enhancing model stability through majority voting. This ensemble approach reduces variance and overfitting while improving generalization, guaranteeing strong performance despite loud or imbalanced data (Pal, 2005).

Gradient Boosting (GB) is implemented to iteratively refine predictions by sequentially training weak learners (e.g., shallow trees) to address residual errors. This method excels in accuracy by focusing on misclassified instances in each iteration, making it highly effective for complex classification challenges (Natekin & Knoll, 2013).

Finally, k-Nearest Neighbors (KNN) is employed as a non-parametric, instance-based learning algorithm. By classifying samples based on proximity to neighboring data points, KNN offers simplicity and interpretability. While computationally lightweight, its performance is contingent on appropriate distance metric selection and feature scaling (Keller et al., 1985).

Risk Label prediction

Based on the trained models, the system predicts the maternal risk category for each patient. The risk levels are classified into three categories: High risk, Mid risk, and Low risk, enabling healthcare professionals to make informed decisions for early intervention.

Performance Evaluation

The efficacy of each classification model is evaluated with important performance measures, incorporating accuracy, precision, recall, and the f1-score, area under the curve (AUC). These metrics provide a comprehensive evaluation of the model's predictive capabilities, ensuring the selection of the most optimal approach for maternal risk classification.

4.Result and Discussion

This section presents the findings of the model evaluation, focusing on feature importance, model performance, and comparing and contrasting various performance measures. The feature importance analysis highlights the most influential factors in predicting pregnancy risk conditions, while the performance of various machine learning models is assessed through accuracy, precision, recall, F1 score, and AUC. Graphical representations, including bar graphs and ROC curves, provide further insights into model efficiency.

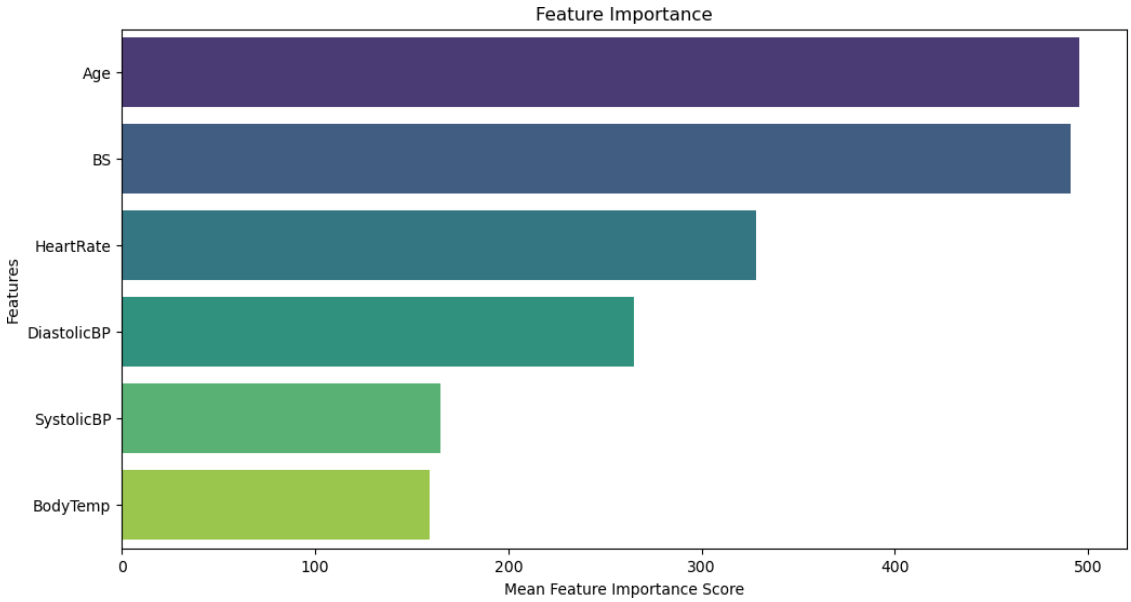


Figure 2: Feature importance

Feature importance analysis using the Random Forest model identifies the most critical features contributing to pregnancy risk classification. As depicted in the bar graph, Age and Blood Sugar (BS) emerge as the most influential features, followed by Heart Rate, Diastolic Blood Pressure, Systolic Blood Pressure, and Body Temperature. These features significantly impact the model's decision-making process, with higher importance scores indicating stronger predictive power.

To classify the dataset for maternal health risk assessment, six machine learning models XGBoost, LightGBM, CatBoost, Random Forest, Gradient Boosting Machines, and KNN were utilized. The dataset comprises six key features relevant to maternal health risk prediction. To ensure the reliability and robustness of the results, cross-validation was conducted. Model performance was

evaluated using various metrics, allowing for a comprehensive comparison of their effectiveness in distinguishing different risk levels (Yildirim et al., 2023; Özbay & Özbay, 2023).

Table 1.2: Cross-Validation Performance of Models.

Model	Accuracy	Precision	Recall	F1 Score	AUC
XGBoost	0.9715	0.9718	0.9715	0.9715	0.9906
LightGBM	0.9703	0.9706	0.9703	0.9703	0.9878
CatBoost	0.9678	0.9679	0.9678	0.9678	0.9921
RandomForest	0.9666	0.9669	0.9666	0.9665	0.9918
GradientBoosting	0.9641	0.9643	0.9641	0.9641	0.9875
KNN	0.9456	0.9457	0.9456	0.9455	0.9789

According to the findings, XGBoost had the best accuracy (97.15%) and F1-score (97.15%), followed closely by LightGBM and CatBoost. The Random Forest model, while slightly behind in accuracy, demonstrates strong performance with an AUC of 0.9918, indicating high discriminatory power between classes.

The bar graph provides a visual representation of model performance metrics, including accuracy, precision, recall, F1-score, and AUC for each model. The results demonstrate that tree-based ensemble models (XGBoost, LightGBM, CatBoost,Random Forest, and Gradient Boosting) outperform KNN in all metrics, highlighting their robustness in handling complex feature

relationships.

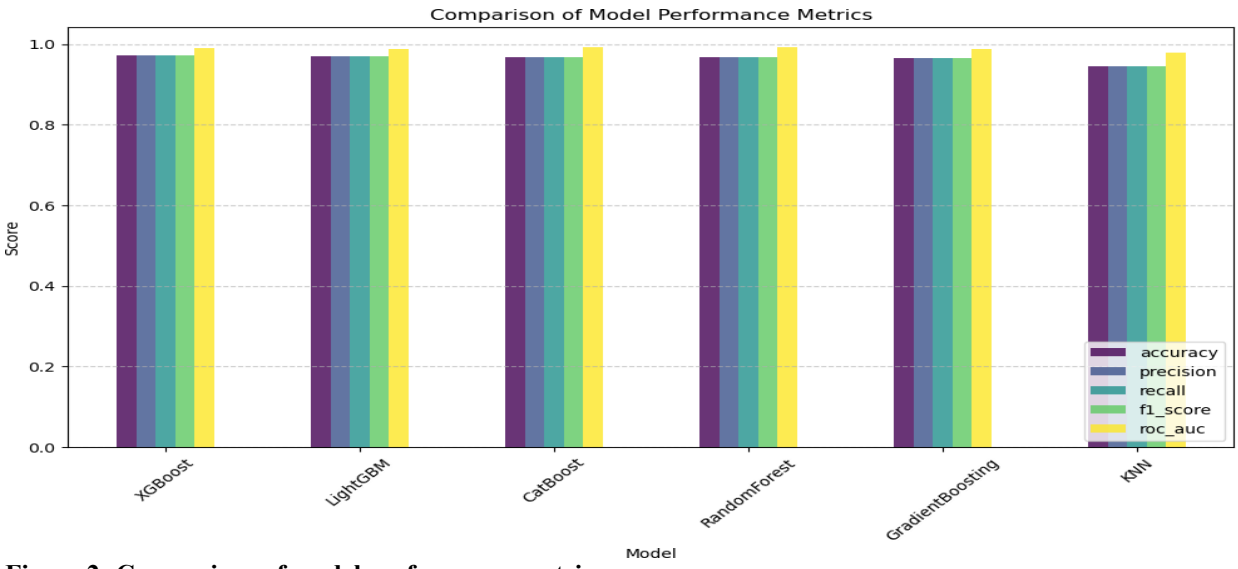


Figure 2: Comparison of model performance metrics

The ROC curve comparison effectively illustrates the trade-off between True Positive Rate (TPR) and False Positive Rate (FPR), validating model performance. CatBoost achieves the highest AUC score (0.9921), followed closely by Random Forest (0.9918) and XGBoost (0.9906), demonstrating strong predictive ability. Gradient Boosting (0.9875) and LightGBM (0.9878) show slightly lower but competitive performance, while KNN records the lowest AUC (0.9789), indicating weaker discriminatory power. This analysis confirms that ensemble-based models outperform KNN, making them more suitable for accurate pregnancy risk classification.

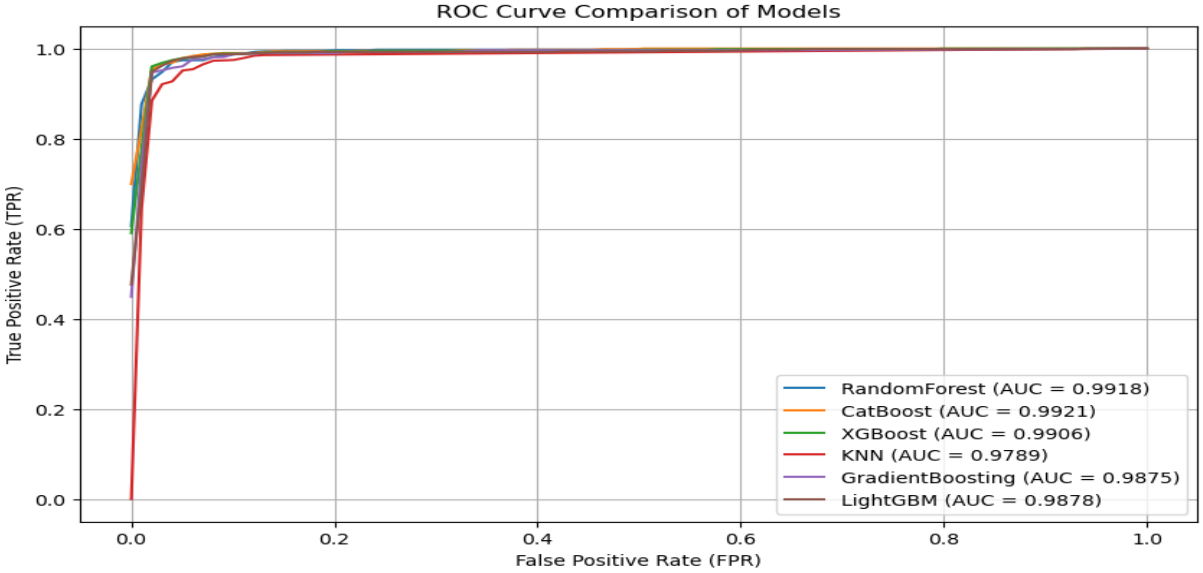


Figure 4: ROC Curve Comparison of Model

5.Conclusion

The comparative analysis demonstrates that ensemble-based models, particularly XGBoost, CatBoost, and Random Forest, achieve superior performance in predicting maternal health risk levels. Feature importance analysis identifies Age and Blood Sugar (BS) as the most influential predictors, emphasizing their significance in maternal health assessment. These findings suggest that leveraging advanced machine learning techniques can significantly enhance early detection and timely intervention for high-risk pregnancies, ultimately reducing maternal morbidity and mortality.

Future research should focus on developing multimodal learning approaches that integrate genetic, lifestyle, and environmental factors alongside clinical parameters to improve prediction accuracy. Additionally, self-learning autonomous AI systems that dynamically adapt to changing maternal health conditions present a complex yet promising area of study. The implementation of federated learning could enable privacy-preserving collaborations among healthcare institutions, facilitating secure and large-scale data sharing. Furthermore, Maternal health risk assessments that are

adaptive and individualized could result from hybrid models that combine deep learning with reinforcement learning, guaranteeing more successful monitoring and intervention tactics.

6.References

Pawar, L., et al. (2022). A robust machine learning predictive model for maternal health risk. In *2022 3rd International Conference on Electronics and Sustainable Communication Systems (ICESC)* (pp. 882–888). IEEE.

Varshavsky, J., et al. (2020). Heightened susceptibility: A review of how pregnancy and chemical exposures influence maternal health. *Reproductive Toxicology, 92*, 14–56.

Umoren, I., et al. (2020). Modeling and prediction of pregnancy risk for efficient birth outcomes using the decision tree classification and regression model.

Ahmed, M., & Kashem, M. A. (2020). IoT-based risk level prediction model for maternal health care in the context of Bangladesh. In *2020 2nd International Conference on Sustainable Technologies for Industry 4.0 (STI)*. IEEE.

Ahmed, M., et al. (2020). Review and analysis of risk factors of maternal health in remote areas using the Internet of Things (IoT). In *InECCE2019: Proceedings of the 5th International Conference on Electrical, Control & Computer Engineering, Kuantan, Pahang, Malaysia, 29th July 2019*. Springer.

Rai, S. K., & Sowmya, K. (2018). A review on the use of machine learning techniques in diagnostic healthcare. *Artificial Intelligent Systems and Machine Learning, 10*(4), 102–107.

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). https://doi.org/10.1145/2939672.2939785

Ke, G., et al. (2017). LightGBM: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems*, 3149–3157.

Hancock, J. T., & Khoshgoftaar, T. M. (2020). CatBoost for big data: An interdisciplinary review. *Journal of Big Data, 7*(1), 1–45. https://doi.org/XXXX

Pal, M. (2005). Random forest classifier for remote sensing classification. *International Journal of Remote Sensing, 26*(1), 217–222.

Natekin, A., & Knoll, A. (2013). Gradient boosting machines: A tutorial. *Frontiers in Neurorobotics, 7*, 21.

Keller, J. M., Gray, M. R., & Givens, J. A. (1985). A fuzzy k-nearest neighbor algorithm.
IEEE Transactions on Systems, Man, and Cybernetics, 15(4), 580–585.

Yildirim, M., et al. (2023). Automatic classification of particles in the urine sediment test with the developed artificial intelligence-based hybrid model. *Diagnostics, 13*(7), 1299.

Özbay, F. A., & Özbay, E. (2023). An NCA-based hybrid CNN model for classification of Alzheimer's disease on Grad-CAM-enhanced brain MRI images. *Turkish Journal of Science and Technology, 18*(1), 139–155.