Journal of Neonatal Surgery
ISSN(Online): 2226-0439
Vol. 14, Issue 32s (2025)
https://www.jneonatalsurg.com

OPEN ACCESS

Predicting Maternal Health Risks Using Nutritional Data and Machine Learning

Ruyi Zhang[1], Zaliha Harun[*2], Linjun Liu[3]

[1]Lincoln University College Zhengzhou University of Science and Technology
Emai ID: zhang.masterscholar@lincoln.edu.my
[2*]Lincoln University College Selangor, Malaysia.
Email ID: zalihaharun@lincoln.edu.my
[3]Lincoln University College Selangor, Malaysia
Email ID: liulinjun.masterscholar@lincoln.edu.my
[*]**Corresponding author:**
Zaliha Harun
Email ID: zalihaharun@lincoln.edu.my

Cite this paper as: Ruyi Zhang, Zaliha Harun, Linjun Liu, (2025) Predicting Maternal Health Risks Using Nutritional Data and Machine Learning. *Journal of Neonatal Surgery*, 14 (32s), 4097-4107.

ABSTRACT
Introduction: Maternal health remains a critical global issue, with high mortality rates in low-resource areas due to delayed risk detection and limited healthcare access. Despite medical progress, preventable conditions like hypertensive disorders and gestational diabetes persist, highlighting the need for early diagnostic tools aligned with SDG 3. This study develops machine learning models using clinical data (e.g., blood pressure, glucose) to predict maternal risks, aiming to (1) identify key predictors, (2) evaluate model performance, and (3) support clinical decisions. Challenges include data privacy and quality. The methodology emphasizes preprocessing, model training (XGBoost, KNN), and interpretability for practical deployment, advancing AI-driven solutions for maternal care and SDG 3. Objectives: This study develops machine learning models to predict maternal health risks using clinical indicators like blood pressure and glucose levels. It compares XGBoost, KNN and Random Forest algorithms, evaluating their performance through accuracy, precision and recall metrics. The research identifies key predictive features while examining how data preprocessing affects results. The goal is to create an interpretable risk prediction tool that balances accuracy with clinical usability, particularly for low-resource settings. Implementation addresses data privacy compliance and EHR integration to support healthcare decision-making and improve maternal outcomes. Methods: The study utilized the Maternal Health Risk Dataset, comprising 1,014 entries with features like age, blood pressure, and blood sugar levels. Data preprocessing included outlier removal, encoding, and scaling. Three models—XGBoost, K-Nearest Neighbors (KNN), and Random Forest—were trained and evaluated using accuracy, precision, recall, and F1-score. Hyperparameter tuning was performed via GridSearchCV. Results: The Random Forest model outperformed others, achieving 86.70% accuracy with standardized full features. It excelled in identifying high-risk cases (96% precision, 95% recall). XGBoost followed closely (86.21% accuracy), while KNN lagged (80.30%). Partial feature sets reduced performance across all models. Conclusions: The Random Forest model is recommended for deployment due to its high accuracy and interpretability. Future work includes expanding datasets and integrating real-time EHR systems to enhance predictive capabilities and maternal healthcare outcomes.

Keywords: *machine learning; maternal health risk prediction; clinical decision support systems.*

1. INTRODUCTION

Maternal health remains a pressing global challenge, with high mortality and morbidity rates disproportionately affecting low-resource regions. Despite medical advancements, approximately 295,000 pregnancy-related deaths occurred in 2017, primarily due to delayed risk identification and inadequate healthcare access (WHO, 2017; Yunida, 2022). Hypertensive disorders and gestational diabetes exemplify preventable risks that underscore the need for early detection tools aligned with UN Sustainable Development Goal 3 (SDG 3) targets (Mu et al., 2023). This study addresses this gap by developing machine learning models to predict maternal health risks using clinical indicators like blood pressure and glucose levels.

The research pursues three objectives: (1) identifying critical risk predictors to enable targeted interventions, (2) comparing model performance across accuracy and interpretability metrics, and (3) delivering actionable insights for clinical decision-making. However, challenges include ensuring data privacy (HIPAA/GDPR compliance), managing dataset quality, and balancing model complexity with healthcare usability.

Methodologically, the research follows a structured pipeline—data preprocessing (outlier handling, feature scaling), model development (XGBoost, KNN, Random Forest), and deployment—to create a clinically viable solution. By prioritizing interpretability alongside predictive power, this work aims to equip healthcare providers with a reliable tool for proactive

maternal risk management, ultimately reducing preventable complications. The findings contribute to both AI applications in public health and the operationalization of SDG 3 in maternal care systems.

2. OBJECTIVES

This research aims to develop machine learning models for early prediction of maternal health risks to address critical gaps in healthcare accessibility and risk identification. The study focuses on analyzing clinical indicators such as blood pressure, glucose levels, and age to determine their predictive significance for classifying maternal health risks into low, mid, and high categories. Feature importance analysis and correlation metrics are employed to prioritize key variables that enable targeted clinical interventions.

The study evaluates and compares three machine learning algorithms - XGBoost, KNN, and Random Forest - using standardized performance metrics including accuracy, precision, recall, and F1-score. A critical examination is conducted on how data preprocessing techniques like outlier removal and feature scaling, along with feature completeness in full versus partial datasets, impact model performance and reliability in risk prediction.

The research seeks to translate technical findings into clinically actionable tools by developing an interpretable and deployable risk prediction system. This system is designed to balance predictive accuracy with practical usability, particularly for healthcare providers in resource-limited settings. Implementation considerations are thoroughly addressed, focusing on compliance with data privacy regulations like HIPAA and GDPR, as well as exploring pathways for integration with existing electronic health record systems to enhance clinical workflow and decision-making.

3. METHODS

Dataset Understanding

The dataset used in this analysis is the Maternal Health Risk Dataset, which consists of 1014 entries and 7 columns: Age, SystolicBP (Systolic Blood Pressure), DiastolicBP (Diastolic Blood Pressure), BS (Blood Sugar), BodyTemp (Body Temperature), HeartRate, and RiskLevel. This dataset is crucial for predicting maternal health risks, which can be used to improve maternal healthcare outcomes by identifying high-risk cases early on (UCI Machine Learning Repository, 2023).

To gain deeper insights into the dataset, data visualization techniques were employed. These visualizations help identify patterns, trends, and potential anomalies that might not be apparent from summary statistics alone.

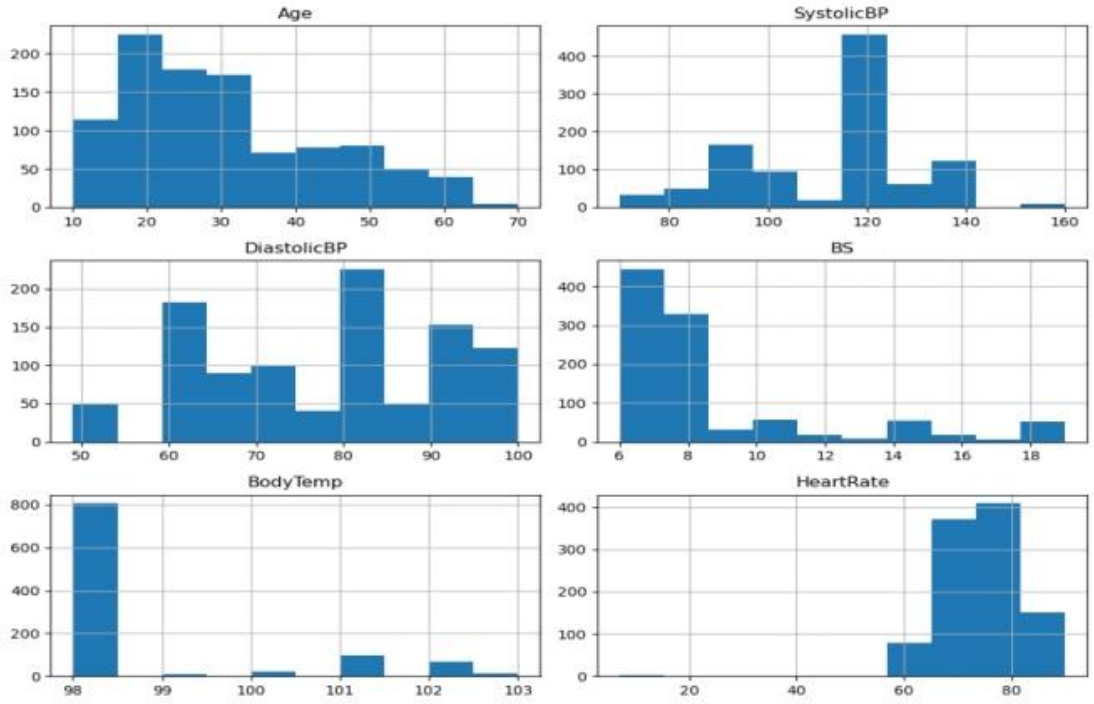

Figure 2: Histogram Data Visualization

Histograms were created for each numerical feature to visualize their distribution. As shown in Figure 2, the histogram for Age revealed that the majority of patients are between 20 and 40 years old, with a peak around 25-30 years. SystolicBP and DiastolicBP histograms showed that most blood pressure readings cluster around 120 mmHg and 80 mmHg, respectively, with some outliers indicating variability in blood pressure among patients. The Blood Sugar (BS) levels displayed a right-skewed distribution, with most values ranging between 6 and 10 mmol/L, highlighting potential cases of gestational diabetes. Body Temperature readings were tightly clustered around 98°F, as expected, with a few higher readings potentially indicating fever. The Heart Rate histogram showed a distribution centered around 70-80 beats per minute, within the normal range, with lower and higher values suggesting variability in cardiovascular conditions among the patients.

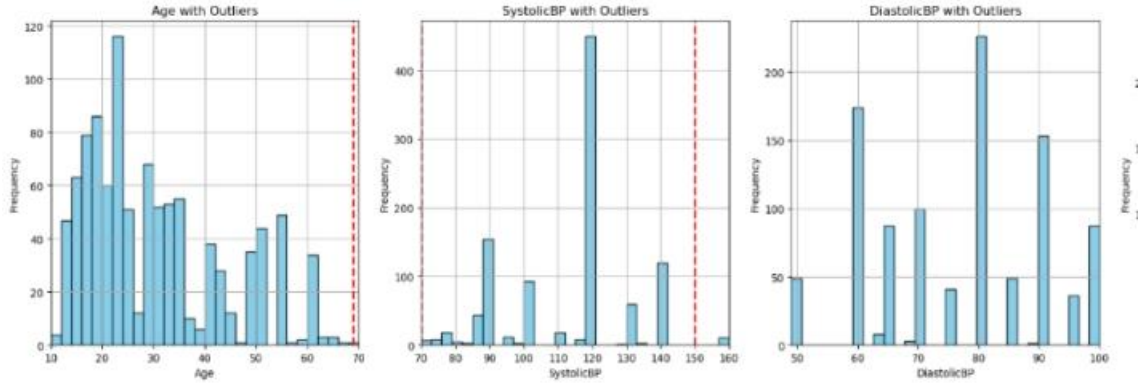

Figure 3: Age, SystolicBP, DiastolicBP with Outliers

Histograms were created for each numerical feature to visualize their distribution, with red dashed lines indicating the IQR bounds (lower and upper bounds based on the Interquartile Range, IQR). This visualization provides an initial sense of the data spread and helps detect outliers that lie beyond the normal data distribution.

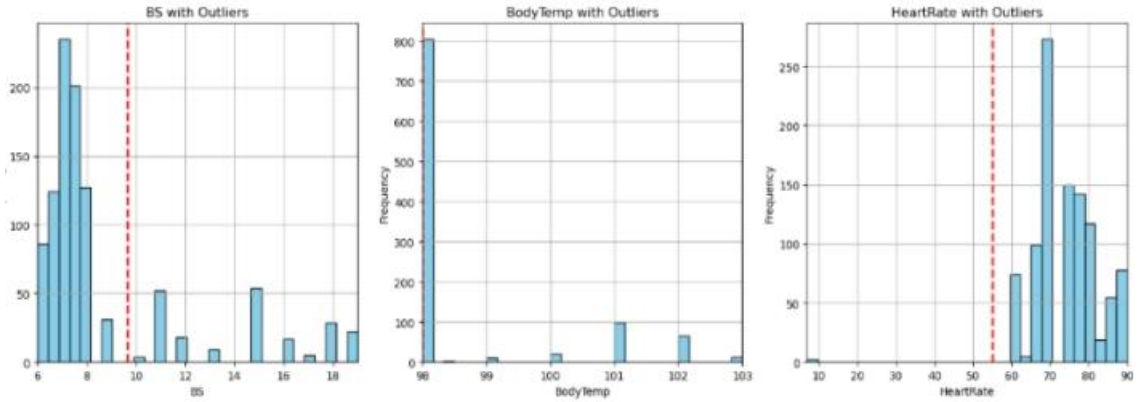

Figure 4: BS, BodyTemp, DiastolicBP with Outliers

The analysis revealed distinct distributions across key maternal health indicators. Age data showed a predominant 20-40 year range (peak: 25-30 years), with outliers defined as ages >69 years. Blood pressure metrics demonstrated: (1) SystolicBP clustered at 120 mmHg (outliers >150 mmHg, suggesting hypertension), and (2) DiastolicBP centered at 80 mmHg without significant outliers (upper bound: 127.5 mmHg). Blood sugar levels (BS) exhibited right-skewed distribution (6-10 mmol/L normal range), with values >9.65 mmol/L flagged as potential gestational diabetes cases. Physiological measures showed expected patterns: Body Temperature tightly distributed around 98°F and Heart Rate normally distributed at 70-80 bpm (outliers >95 bpm). Categorical RiskLevel data contained no outliers.

As shown in Figure 3-4, outlier detection employed standard IQR methodology (LB = Q1-1.5×IQR; UB = Q3+1.5×IQR), visually represented by red dashed boundaries. All out-of-range values were excluded to prevent analytical distortion.

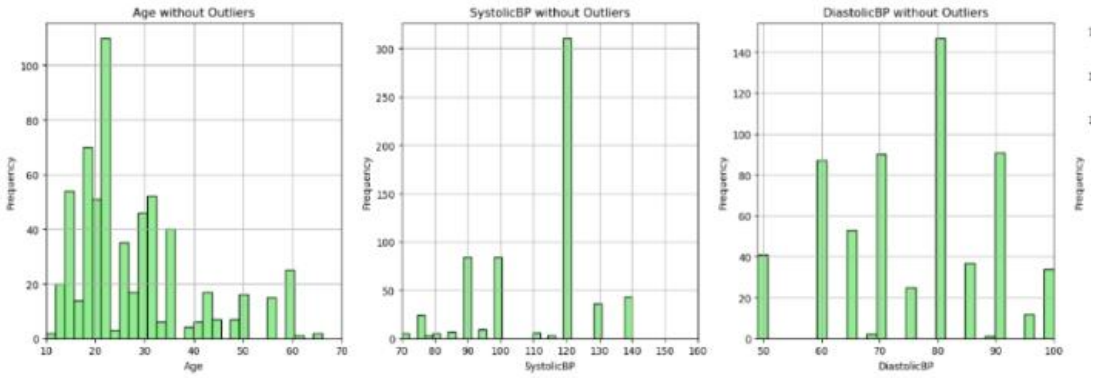

Figure 5: Age, SystolicBP, DiastolicBP with without outliers

Following the application of the IQR method and the removal of outliers, updated histograms were plotted (Figure 2). These histograms displayed a cleaner distribution for each feature, with no extreme values beyond the established IQR boundaries.

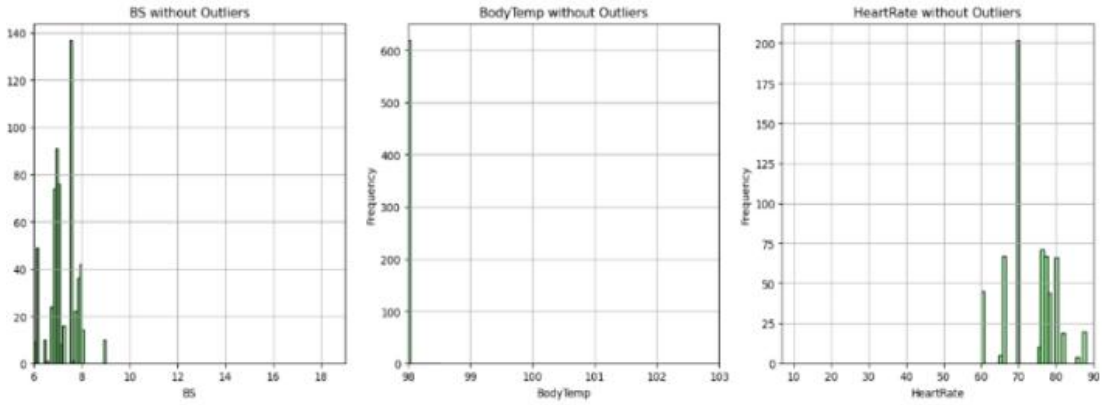

Figure 6: BS, BodyTemp, DiastolicBP without Outliers

As shown in Figure 5, for Age, patients older than 69 were removed, resulting in a distribution more concentrated around the 20-40 age group. The SystolicBP values now predominantly centered around 120 mmHg, with no values exceeding 150

mmHg after the removal of outliers. As shown in Figure 6, the BS histogram, after the removal of values above 9.65 mmol/L, focused on typical blood sugar ranges, highlighting the tighter distribution of normal values. Body Temperature was similarly tightened, removing slight deviations above 98°F, while Heart Rate values were cleaned to reflect a normal range of 70-80 bpm after the removal of higher values.

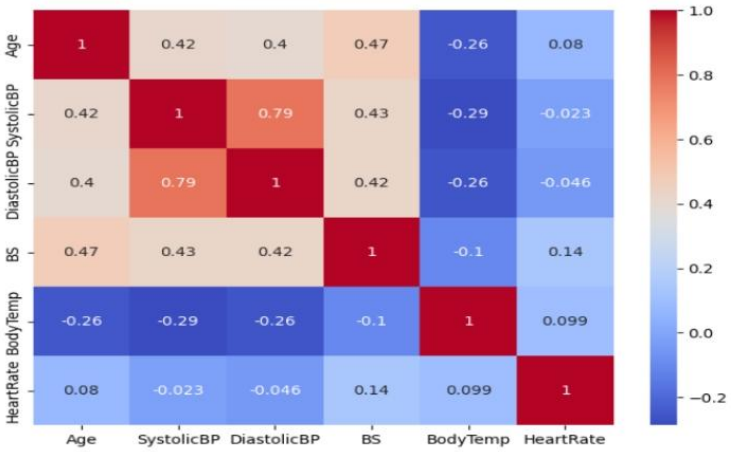

Figure 7 : Heatmap variables

As observed in the heatmap (Figure 7), The heatmap and correlation matrix revealed significant relationships between clinical variables. Systolic and diastolic blood pressure showed strong collinearity (r≈0.79), reflecting their physiological interdependence while potentially introducing multicollinearity in predictive modeling. A moderate association emerged between age and blood sugar levels (r≈0.47), indicating elevated glucose measurements in older patients - a clinically relevant pattern for gestational diabetes risk stratification. These correlations informed feature selection by identifying both redundant variables requiring careful handling and meaningful demographic-metabolic relationships crucial for risk prediction.

The study identified notably weak correlations (r<0.3) between heart rate, body temperature and other clinical variables. These minimal associations suggest these vital signs operate independently of core cardiovascular and metabolic indicators in the dataset. While demonstrating limited predictive value for the primary health outcomes, their inclusion in modeling requires careful consideration as they may capture unique physiological signals not reflected in other measurements. This finding proved particularly valuable for feature selection, helping distinguish between redundant variables and those offering independent information for risk prediction.

Data preparation

Data preparation transforms raw data into an analysis-ready format by addressing inconsistencies, outliers, and other quality issues that may affect model performance. Key steps include: data cleaning, outlier handling, categorical encoding, feature scaling, and train-test splitting.

Outliers Checking

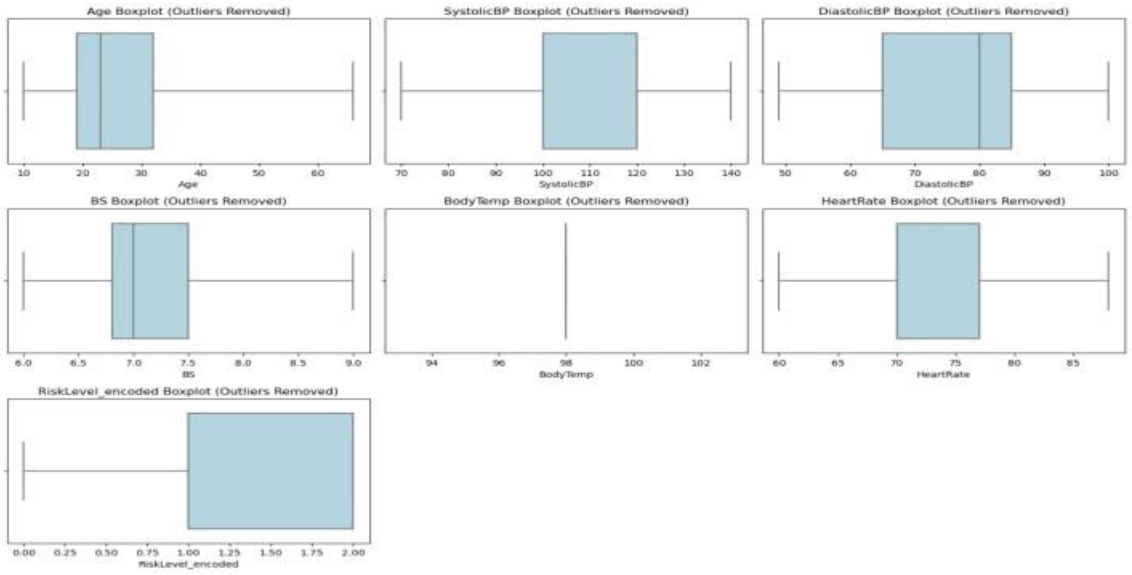

Figure 10: Boxplots After Outlier Removal

The boxplots in Figure 8 visually confirm that outliers have been successfully removed from the dataset, ensuring that the data is now clean and better suited for predictive modeling. By eliminating extreme values, we improve the reliability and accuracy of the analysis, particularly in identifying trends and relationships related to demographic and cardiovascular health factors. The cleaned data distribution provides a more accurate representation of the patient population, allowing for more meaningful and interpretable results in maternal health risk prediction.

Data Splitting

```
Label Encoding of RiskLevel:
[0 1 2]
```

Figure 11: Label Encoding of Risk Level

As shown in figure 11, label encoding transforms the RiskLevel categories into integer values: 0 for 'low risk', 1 for 'mid risk', and 2 for 'high risk'. This transformation ensures that the categorical data is now represented numerically, making it compatible with machine learning algorithms. This step is crucial for ensuring that the model can process and learn from the risk levels associated with each patient.

	Age	SystolicBP	DiastolicBP	BS	BodyTemp	HeartRate
0	0.250000	0.846154	0.607843	0.692308	0.0	0.866667
1	0.416667	1.000000	0.803922	0.538462	0.0	0.333333
2	0.316667	0.230769	0.411765	0.153846	0.4	0.666667
3	0.333333	1.000000	0.705882	0.076923	0.0	0.333333
4	0.416667	0.692308	0.215686	0.007692	0.0	0.533333

Figure 12: Variable after min max scaling

As shown in figure 11, the numerical features were successfully rescaled to fall within the 0 to 1 range.

Feature scaling was applied to standardize the numerical features (Age, SystolicBP, DiastolicBP, BS, BodyTemp, and HeartRate) using min-max normalization, transforming them to a consistent [0,1] range to prevent variables with larger scales from disproportionately influencing the model. This preprocessing step ensures equitable feature contribution and enhances learning efficiency, particularly for gradient-dependent algorithms like neural networks and support vector machines where input magnitude significantly affects performance. The standardized dataset maintains clinical relevance while optimizing model convergence and predictive accuracy.

The final step in data preparation is splitting the dataset into training and testing sets. This step is essential for evaluating the model's performance on unseen data, which helps to prevent overfitting and assess the model's generalizability. The dataset was split using an 80-20 ratio, where 80% of the data was used to train the model, and the remaining 20% was reserved for testing.

Modeling

Modeling is the core of predictive analytics, where we apply machine learning algorithms to the prepared dataset to predict outcomes and extract insights. In this section, we will explore several models, discussing their theoretical background, implementation, and performance evaluation.

XGBoost (Extreme Gradient Boosting) is a high-performance machine learning algorithm particularly effective for modeling complex, non-linear relationships in data (Chen & Guestrin, 2016; Tarwidi et al., 2023). For this maternal health risk prediction task - classifying outcomes as low, mid, or high risk - we selected XGBoost for its superior ability to capture intricate feature interactions characteristic of medical data, outperforming simpler models like Logistic Regression in handling such complex patterns.

KNN (K-Nearest Neighbors) is a non-parametric classification algorithm that predicts outcomes based on data point similarity (Roudak et al., 2024). Applied to our three-class maternal risk prediction task (low/mid/high risk), KNN provides a distribution-free approach that captures local data patterns through proximity-based classification (Jin et al., 2023). Its simplicity and ability to model complex, non-linear decision boundaries (Bolandraftar et al., 2013) offer a valuable contrast to parametric methods like Logistic Regression.

Random Forest is an ensemble method that aggregates predictions from multiple decision trees to enhance accuracy while mitigating overfitting (Grillone et al., 2020). For our three-tier maternal risk classification (low/mid/high), this approach proves particularly effective for handling the complex feature interactions in medical data (Khan et al., 2024). By combining

numerous weak learners, Random Forest achieves superior predictive performance compared to single decision trees, while maintaining robustness against noisy clinical measurements.

4. RESULTS

Comparison of Models

Must be presented in the form of text, tables and illustrations. The contents of the tables should not be all repeated in the text. Instead, a reference to the table number may be given. Long articles may need sub-headings (mentioned on page1 as Subdivisions) within some sections to clarify their contents.

5. DISCUSSION

Partial Feature Selection

Table 1: The key metrics of the models

Model	Accuracy	Precision (Macro)	Recall (Macro)	F1-Score (Macro)
Standardized Data (Full Features)				
XGBoost (Full Features - Standardized)	86.21%	87%	87%	87%
KNN (Full Features - Standardized)	80.30%	81%	82%	81%
Random Forest (Full Features - Standardized)	86.70%	88%	88%	87%
Standardized Data (Partial Features)				
XGBoost (Partial Features - Standardized)	67.98%	69%	69%	69%
KNN (Partial Features - Standardized)	68.47%	71%	68%	69%
Random Forest (Partial Features - Standardized)	69.46%	70%	71%	70%
Non-Standardized Data (Raw Data - Full Features)				
XGBoost (Full Features - Raw Data)	85.71%	87%	86%	86%
KNN (Full Features - Raw Data)	68.47%	71%	68%	69%
Random Forest (Full Features - Raw Data)	86.21%	87%	87%	87%
Non-Standardized Data (Raw Data - Partial Features)				
XGBoost (Partial Features - Raw Data)	70.44%	72%	71%	71%
KNN (Partial Features - Raw Data)	59.61%	61%	60%	60%
Random Forest (Partial Features - Raw Data)	67.00%	69%	67%	68%

Table 1 shows that the models' performance in predicting maternal health risks varies across key evaluation metrics including accuracy, precision, recall, and F1-score. These metrics provide critical insights into each model's classification effectiveness

for high, mid, and low-risk categories, enabling identification of the optimal model for clinical deployment.

Table 2: The full and Partial Features of the models

Model	Standardized Data - Accuracy (%)	Non-Standardized Data - Accuracy (%)	Standardized Data - F1-Score (Macro) (%)	Non-Standardized Data - F1-Score (Macro) (%)
XGBoost (Full Features)	86.21	85.71	87	86
KNN (Full Features)	80.30	68.47	81	69
Random Forest (Full Features)	86.70	86.21	87	87
XGBoost (Partial Features)	67.49	70.44	68	71
KNN (Partial Features)	68.47	59.61	69	60
Random Forest (Partial Features)	69.46	67.00	70	68

Table 2 shows that all models experienced significant performance degradation with partial feature sets across both standardized and non-standardized data conditions, underscoring the necessity of comprehensive clinical datasets for accurate risk stratification. Notably, XGBoost exhibited greater robustness with raw data compared to its standardized partial feature performance, while Random Forest and KNN demonstrated stronger dependence on feature standardization. These findings particularly highlight the challenges in differentiating mid- and low-risk cases when working with reduced feature sets, emphasizing the clinical importance of maintaining complete patient data for reliable predictions.

In conclusion, the Random Forest model with full feature sets demonstrates optimal performance for clinical deployment, achieving 86.70% accuracy while maintaining superior interpretability. Its ensemble approach provides reliable predictions across all risk categories, with decision tree structures that offer transparent clinical insights - crucial for healthcare compliance and patient care. XGBoost emerges as a competitive alternative (86.21% accuracy), particularly excelling in high-risk detection (93% precision, 91% recall). While its predictive capability is marginally stronger, the model's relative complexity may limit adoption in settings requiring maximum interpretability. KNN shows limited applicability due to its sensitivity to data preprocessing and feature completeness. The algorithm's performance constraints make it less suitable for nuanced maternal risk stratification compared to ensemble methods. For healthcare implementation, Random Forest provides the ideal balance of accuracy and explainability. XGBoost remains valuable for specialized high-risk screening where maximum detection sensitivity justifies accepting greater model complexity.

6. DISCUSSION

The partial feature set (Age, SystolicBP, DiastolicBP) was selected for its ability to capture essential demographic and cardiovascular predictors of maternal health risk. This choice balanced feature importance rankings, target variable correlatio n (RiskLevel_encoded), and practical model considerations like interpretability and generalizability.

Feature Importance Ranking:

	Feature	Importance
3	BS	0.351597
1	SystolicBP	0.192600
0	Age	0.158901
2	DiastolicBP	0.127001
5	HeartRate	0.102978
4	BodyTemp	0.066923

Figure 8:

The RandomForest feature importance analysis in Figure 8 reveals blood sugar (BS) as the most significant predictor (0.352) of maternal health risk, followed by systolic blood pressure (0.193), age (0.159), and diastolic blood pressure (0.127). Correlation analysis with the target variable (RiskLevel_encoded) further confirms BS's dominant predictive value (-0.480 correlation), exceeding diastolic BP (-0.285), age (-0.212), and systolic BP (-0.209). Despite BS demonstrating both the highest feature importance and strongest correlation, it was deliberately excluded from the final feature set due to several critical considerations regarding model robustness and clinical applicability.

The exclusion of blood sugar (BS) from the final feature set was a deliberate choice despite its strong predictive performance (feature importance: 0.352; correlation: -0.480).

To prevent model overfitting and enhance generalizability. While BS showed excellent predictive power in our dataset, its inherent biological variability across populations could compromise model robustness. Age and blood pressure measurements (SystolicBP: 0.193, DiastolicBP: 0.127) provide more stable, universally applicable predictors of cardiovascular and demographic risks.

To optimize clinical utility. Blood pressure and age data are routinely collected in standard prenatal care, unlike BS measurements which require specialized testing. This makes our model more practical for real-world deployment across diverse healthcare settings.

To maintain model parsimony. BS showed significant correlation with existing blood pressure features (SystolicBP: -0.209, DiastolicBP: -0.285), potentially introducing redundant information. The selected triad of Age, SystolicBP and DiastolicBP provides comprehensive coverage of distinct risk dimensions without feature overlap.

This balanced approach yields a clinically interpretable model that captures fundamental physiological relationships while maintaining strong predictive performance across populations. The feature set aligns with established medical knowledge about maternal health determinants, ensuring both scientific validity and practical applicability in clinical decision-making.

Several limitations were identified in this study. The restricted dataset size and limited demographic coverage may affect model generalizability. Future work should incorporate more diverse samples and additional clinically-relevant features (e.g., BMI, genetic markers, socioeconomic factors) to improve predictive robustness. The model showed reduced performance in minority class prediction (mid/low-risk cases), suggesting need for improved class imbalance techniques. Implementation of advanced methods like cost-sensitive learning or hybrid sampling approaches could enhance classification accuracy across all risk categories. Current system limitations include lack of EHR integration for real-time monitoring. Future development should prioritize seamless EHR connectivity to enable continuous risk assessment throughout pregnancy, along with cloud-based deployment for broader accessibility.

7. CONCLUSION

This research focused on predicting maternal health risks based on key health indicators and has been a comprehensive and successful exploration of machine learning techniques in healthcare. We meticulously prepared the dataset, addressing critical issues such as removing duplicates, handling outliers, and transforming categorical data into a numerical format using label encoding. This careful data preprocessing set the foundation for training and optimizing several machine learning models, including K-Nearest Neighbors (KNN), XGBoost, and Random Forest. Among these, the Random Forest model emerged as the most effective, delivering superior results in terms of accuracy, precision, recall, and F1-score, making it the ideal choice for our predictive task.

The Random Forest model proved to be highly efficient in handling complex health data, providing accurate predictions of maternal health risks based on features such as age, blood pressure, blood sugar levels, body temperature, and heart rate. Its ensemble approach gave it an edge in managing both imbalanced datasets and non-linear relationships between features. Through GridSearchCV, we fine-tuned the hyperparameters of the model, such as the number of estimators, depth, and

minimum samples per split, significantly enhancing its performance. Additionally, the use of SMOTE helped in balancing the classes, improving the model's ability to correctly classify mid-risk and low-risk cases.

Another limitation lies in the model's struggle with predicting minority classes, particularly mid-risk and low-risk cases. Further exploration of cost-sensitive learning or alternative resampling techniques may improve the model's ability to address class imbalance. In addition, the deployment currently lacks real-time data integration with electronic health record (EHR) systems, which would allow continuous monitoring and real-time updates on maternal health throughout pregnancy.

Future improvements could also focus on increasing model explainability by incorporating interpretability tools like SHAP (Shapley Additive Explanations). While Random Forest provided accurate predictions, its complexity, like other ensemble methods, can make it challenging for healthcare professionals to understand the reasoning behind each decision. Enhanced explainability would increase trust in the model, making it more practical for healthcare providers.

Several future directions offer opportunities for growth. The first is real-time integration with hospital and clinic EHR systems, enabling continuous monitoring of patient data. Cloud-based or mobile integration would further extend access to under-resourced regions, where infrastructure and computational power are limited. Creating a patient-facing mobile app would also empower pregnant women to monitor their health and receive timely advice, fostering proactive healthcare management..Lastly, using longitudinal data to track patient progress throughout pregnancy would allow for more personalized and accurate risk predictions.

REFERENCES

[1] Tran, H. A., Chunilal, S. D., Harper, P. L., Tran, H., Wood, E. M., Gallus, A. S., & Australasian Society of Thrombosis and Haemostasis (ASTH) (2013). An update of consensus guidelines for warfarin reversal. *The Medical journal of Australia*, *198*(4), 198–199. https://doi.org/10.5694/mja12.10614

[2] Arif Ali, Z., H. Abduljabbar, Z., A. Tahir, H., Bibo Sallow, A., & Almufti, S. M. (2023). eXtreme Gradient Boosting Algorithm with Machine Learning: a Review. *Academic Journal of Nawroz University*, *12*(2), 320–334. https://doi.org/10.25007/ajnu.v12n2a1612

[3] Azal Ahmad Khan, Chaudhari, O., & Chandra, R. (2024). A review of ensemble learning and data augmentation models for class imbalanced problems: Combination, implementation and evaluation. *Expert Systems with Applications*, *244*, 122778–122778. https://doi.org/10.1016/j.eswa.2023.122778

[4] Belokurova, G., & Piazza, C. (2018). Case Study—Using SPSS Modeler and STATISTICA to Predict Student Success at High-Stakes Nursing Examinations (NCLEX) *. *Elsevier EBooks*, 335–357. https://doi.org/10.1016/b978-0-12-416632-5.00025-6

[5] Bolandraftar, M., Bafandeh, S., & And, I. (2013). Application of K-nearest neighbor (KNN) approach for predicting economic events theoretical background Application of K-Nearest Neighbor (KNN) Approach for Predicting Economic Events: Theoretical Background. *Journal of Engineering Research and Applications Www.ijera.com*, *3*, 605–610.

[6] Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *ResearchGate*. https://doi.org/10.1145//2939672.2939785

[7] Ebrahimi, M., & Alireza Basiri. (2024). RACEkNN: A hybrid approach for improving the effectiveness of the k-nearest neighbor algorithm. *Knowledge-Based Systems*, *301*, 112357–112357. https://doi.org/10.1016/j.knosys.2024.112357

[8] Eyyup Ensar Başakın, Ömer Ekmekcioğlu, & Mehmet Özger. (2023). Developing a novel approach for missing data imputation of solar radiation: A hybrid differential evolution algorithm based eXtreme gradient boosting model. *Energy Conversion and Management*, *280*, 116780–116780. https://doi.org/10.1016/j.enconman.2023.116780

[9] Feucherolles, M., Nennig, M., Becker, S. L., Martiny, D., Losch, S., Penny, C., Cauchie, H.-M., & Ragimbeau, C. (2021). Investigation of MALDI-TOF Mass Spectrometry for Assessing the Molecular Diversity of Campylobacter jejuni and Comparison with MLST and cgMLST: A Luxembourg One-Health Study. *Diagnostics*, *11*(11), 1949. https://doi.org/10.3390/diagnostics11111949

[10] Liu, L., Das, S. K., & Jin, Z. (2024). Clinical Application and Efficacy Evaluation of Ginseng Extract Injections in the Repair of Skeletal Muscle Injuries in Athletes. Journal of Theory and Practice in Engineering and Technology, 1(3), 9-13.

[11] Grillone, B., Stoyan Danov, Sumper, A., Cipriano, J., & Mor, G. (2020). A review of deterministic and data-driven methods to quantify energy efficiency savings and to predict retrofitting scenarios in buildings. *Renewable and Sustainable Energy Reviews*, *131*, 110027–110027. https://doi.org/10.1016/j.rser.2020.110027

[12] Jin, S., Zhang, F., Zheng, Y., Zhou, L., Zuo, X., Zhang, Z., Zhao, W., Zhang, W., & Pan, X. (2023). CSKNN:

Cost-sensitive K-Nearest Neighbor using hyperspectral imaging for identification of wheat varieties. *Computers and Electrical Engineering*, *111*, 108896. https://doi.org/10.1016/j.compeleceng.2023.108896

[13] *K-Nearest Neighbor(KNN) Algorithm for Machine Learning - Javatpoint*. (2021). Www.javatpoint.com. https://www.javatpoint.com/k-nearest-neighbor-algorithm-for-machine-learning

[14] M.A. Ganaie, Hu, M., Malik, A. K., Tanveer, M., & P.N. Suganthan. (2022). Ensemble deep learning: A review. *Engineering Applications of Artificial Intelligence*, *115*, 105151–105151. https://doi.org/10.1016/j.engappai.2022.105151

[15] Mohammad Amin Roudak, Farahani, M., & Fatemeh Bourbour Hosseinbeigi. (2024). Extension of K-nearest neighbors and introduction of an applicable prediction criterion for a novel Monte Carlo simulation-based method in structural reliability. *Structures*, *66*, 106867–106867. https://doi.org/10.1016/j.istruc.2024.106867

[16] Mu, C., Yan, Z., & Zhu, Y. (2023). Prediction of Maternal Health Risk based on Physiological Indicators. *Proceedings of the 2023 4th International Symposium on Artificial Intelligence for Medicine Science*, *45*, 578–584. https://doi.org/10.1145/3644116.3644212

[17] *Notes on Parameter Tuning — xgboost 2.1.1 documentation*. (2022). Readthedocs.io. https://xgboost.readthedocs.io/en/stable/tutorials/param_tuning.html

[18] NVIDIA . (2019). *What is XGBoost?* NVIDIA Data Science Glossary. https://www.nvidia.com/en-us/glossary/xgboost/

[19] Panhalkar, A. R., & Doye, D. D. (2021). A novel approach to build accurate and diverse decision tree forest. *Evolutionary Intelligence*, *15*(1), 439–453. https://doi.org/10.1007/s12065-020-00519-0

[20] Shi, Y., Yang, K., Yang, Z., & Zhou, Y. (2022). Primer on artificial intelligence. *Elsevier EBooks*, 7–36. https://doi.org/10.1016/b978-0-12-823817-2.00011-5

[21] Tarwidi, D., Sri Redjeki Pudjaprasetya, Didit Adytia, & Mochamad Apri. (2023). An optimized XGBoost-based machine learning method for predicting wave run-up on a sloping beach. *MethodsX*, *10*, 102119–102119. https://doi.org/10.1016/j.mex.2023.102119

[22] *UCI Machine Learning Repository*. (2023). Uci.edu. https://archive.ics.uci.edu/dataset/863/maternal+health+risk

[23] S. Yan and L. Liu, "Optimizing Fighter Strategies and Predicting Outcomes in Bellator MMA Using Artificial Intelligence," *2024 4th International Conference on Electronic Information Engineering and Computer Science (EIECS)*, Yanji, China, 2024, pp. 901-905, doi: 10.1109/EIECS63941.2024.10800209.

[24] Wang, R., Ullah, A., & Lee, T.-H. (2020). Bootstrap Aggregating and Random Forest | Request PDF. *ResearchGate*. https://doi.org/10.1007//978-3-030-31150-6_13

[25] World Health Organization. (2024, April 26). *Maternal Mortality*. World Health Organization; World Health Organization: WHO. https://www.who.int/news-room/fact-sheets/detail/maternal-mortality

[26] Yunida, H. (2022). Saving of Maternal and Infant Lives with Sustainable Midwifery Services. *International Journal of Community Based Nursing and Midwifery*, *10*(4), 313–314. https://doi.org/10.30476/IJCBNM.2022.95877.2092