AN AI-BASED MATERNAL HEALTH RISK PREDICTION MODEL USING MULTILAYER PERCEPTRON AND COMPREHENSIVE DATA PREPROCESSING

[1]Dien Kartika Lutfi
[1] Prodi S1 Informatika Universitas Sebelas Maret, PSDKU Kebumen
Email: dienkartika24@gmail.com

ABSTRACT

Background: Maternal health risk detection is essential for reducing pregnancy-related complications and improving maternal and fetal outcomes. However, conventional risk assessment remains challenging due to data quality issues and limitations in accurately identifying high-risk pregnancies. Artificial Intelligence (AI) offers promising solutions for improving maternal health risk prediction. **Objectives:** This study aims to develop and evaluate an AI-based maternal health risk prediction model using the Multilayer Perceptron (MLP) classifier and to assess the contribution of comprehensive data preprocessing techniques in improving prediction performance. **Methods:** The study utilized the Maternal Health and High-Risk Pregnancy Dataset containing maternal clinical and physiological indicators. Data preprocessing included missing value imputation, duplicate removal, outlier treatment using the capping method, feature scaling with Robust Scaler, and class balancing through Random Oversampling (ROS). The MLP classifier was trained using a supervised learning approach and evaluated using Accuracy, Precision, Recall, F1-score, Specificity, and Classification Error. **Results:** The proposed model achieved an accuracy of 97.60%, precision of 97.07%, recall of 98.20%, F1-score of 97.62%, specificity of 96.99%, and a classification error of 2.40%, demonstrating excellent predictive performance. **Conclusions:** The integration of comprehensive preprocessing techniques with the MLP classifier significantly improves maternal health risk prediction. The proposed model shows strong potential to support early risk detection, clinical decision-making, and preventive maternal healthcare through accurate and reliable AI-based prediction.

Keywords: Artificial Intelligence, Maternal Health, Risk Prediction, Multilayer Perceptron, Machine Learning.

INTRODUCTION

Maternal health remains a critical global public health concern, particularly in developing and middle-income countries where maternal morbidity and mortality continue to contribute substantially to adverse health outcomes. According to the World Health Organization (WHO), preventable maternal deaths continue to occur due to complications during pregnancy and childbirth, including hypertensive disorders, gestational diabetes, hemorrhage, infections, and cardiovascular complications. These conditions may significantly increase maternal and fetal morbidity when they are not detected and managed appropriately. In addition, recent clinical evidence has highlighted that endocrine, metabolic, inflammatory, and physiological changes during pregnancy are closely associated with maternal and neonatal outcomes, emphasizing the importance of continuous maternal health monitoring and early risk assessment (WHO, 2025)(Li et al., 2024).

Despite continuous improvements in maternal healthcare services, maternal risk assessment in many healthcare settings still relies primarily on conventional screening methods and clinical judgment. These approaches often require extensive clinical expertise and may not adequately identify complex interactions among multiple maternal risk factors, particularly in resource-limited healthcare environments. Consequently, delayed identification of high-risk pregnancies may reduce opportunities for timely intervention and increase the likelihood of adverse maternal and neonatal outcomes. The increasing availability of electronic medical records and routinely collected maternal clinical data has encouraged the development of Artificial Intelligence (AI)-based decision-support systems capable of assisting healthcare professionals in predicting maternal health risks more efficiently.

Recent developments in Artificial Intelligence (AI) and Machine Learning (ML) have demonstrated considerable potential for improving maternal health risk prediction. In a systematic review, (Islam et al., 2025) reported that AI-based risk assessment tools consistently improve prediction performance across maternal, reproductive, and mental healthcare by analyzing routinely collected healthcare data and supporting clinical decision-making. Similarly, (Malik et al., 2024) reviewed machine learning applications for preeclampsia prediction and concluded that AI algorithms have shown promising capability for identifying women at risk of pregnancy-related complications. Another systematic review by (Özcan & Peker, 2025) also emphasized that machine learning has become an increasingly effective approach for predicting preeclampsia while highlighting the importance of appropriate model validation and data preprocessing.

Several recent studies have successfully developed machine learning models for maternal health risk prediction. (Al Mashrafi et al., 2024) investigated multiple machine learning algorithms for maternal risk classification using maternal physiological indicators and demonstrated that AI-based models can effectively classify maternal health risk levels. (Khadidos et al., 2024) proposed an ensemble machine learning framework that improved predictive performance by integrating multiple classification models.

(Xinyu Pi, 2021) compared different machine learning algorithms for high-risk pregnancy prediction and demonstrated the effectiveness of advanced machine learning techniques for maternal risk classification. In addition, (Malde et al., 2025) developed a machine learning framework utilizing sparse maternal clinical information and vital signs, demonstrating that reliable maternal risk prediction can be achieved even in healthcare settings with limited data availability. Furthermore, (Mamun et al., 2025) introduced an explainable machine learning framework that integrates feature optimization and Explainable Artificial Intelligence (XAI), enabling improved prediction accuracy while increasing model transparency and interpretability for healthcare professionals. More recently, (Pavagada & Vemuri, 2025) compared several machine learning algorithms for maternal health risk prediction and confirmed that machine learning techniques provide reliable classification performance for identifying maternal health risk categories.

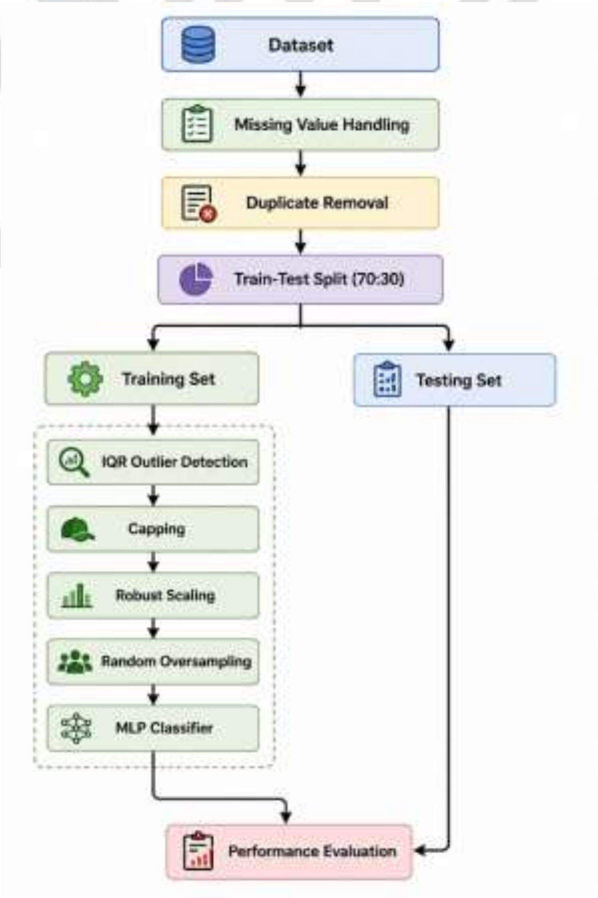
The predictive performance of machine learning models, however, is strongly influenced by the quality of the input data. Maternal health datasets generally contain heterogeneous clinical and physiological variables, including maternal age, blood pressure, blood glucose level, body mass index, heart rate, previous pregnancy complications, diabetes history, and other health indicators. Nevertheless, healthcare datasets frequently contain missing values, duplicate records, outliers, inconsistent feature scales, and imbalanced class distributions that may substantially reduce predictive performance if not properly addressed. Previous studies have consistently emphasized that preprocessing techniques including missing value imputation, outlier treatment, feature normalization, feature selection, and class balancing, play a crucial role in improving model robustness and classification accuracy. Recent studies published in (Gao et al., 2025), (Vasudevan et al., 2025) and (Tzimourta et al., 2025) further highlighted that appropriate preprocessing strategies and high quality clinical data are essential for developing reliable AI based maternal health prediction systems.

Although previous studies have demonstrated promising results for maternal health risk prediction, several research gaps remain. Most existing studies have primarily focused on comparing machine learning algorithms such as Random Forest, Support Vector Machine (SVM), Decision Tree, XGBoost, and ensemble learning methods. Comparatively fewer studies have investigated the combined contribution of comprehensive preprocessing techniques prior to neural network-based classification. In particular, limited research has evaluated the integration of missing value imputation, duplicate removal, outlier treatment using the capping method, Robust Scaler normalization, and Random Oversampling (ROS) within a unified Multilayer Perceptron (MLP) framework. Consequently, the effectiveness of combining these preprocessing strategies with MLP for maternal health risk prediction remains insufficiently explored.

Therefore, this study proposes an Artificial Intelligence based maternal health risk prediction model using the Multilayer Perceptron (MLP) classifier integrated with comprehensive data preprocessing techniques. The proposed preprocessing pipeline includes missing value imputation, duplicate removal, outlier treatment using the capping method, Robust Scaler normalization, and Random Oversampling (ROS) to improve data quality before model development. The proposed model is evaluated to determine its effectiveness in predicting maternal health risk levels and to investigate the contribution of comprehensive preprocessing techniques toward improving predictive performance. The findings are expected to contribute to the development of intelligent clinical decision-support systems capable of supporting early maternal risk detection and preventive maternal healthcare.

METHODS

This study utilized the Maternal Health and High-Risk Pregnancy Dataset obtained from Kaggle (Chayan, 2024). The dataset contains 1,205 maternal health records and includes various clinical and physiological attributes associated with pregnancy risk assessment, namely Age, Systolic Blood Pressure, Diastolic Blood Pressure, Blood Sugar (BS), Body Temperature, Body Mass Index (BMI), Previous Complications, Preexisting Diabetes, Gestational Diabetes, Mental Health Status, and Heart Rate. The target variable is Risk Level, which classifies maternal conditions into different risk categories.



Several preprocessing techniques were applied to improve data quality. Missing values in numerical attributes were handled using linear interpolation, while missing values in the target variable were imputed using the mode. Duplicate records were subsequently removed, reducing the dataset from 1,205 to 1,187 instances.

The cleaned dataset was then divided into 70% training data and 30% testing data. Outliers in the training data were identified using the Interquartile Range (IQR) method and treated using the capping technique to minimize the influence of extreme values without removing observations. To address variations in feature ranges and improve model stability, data normalization was performed using the Robust Scaler. Furthermore, class imbalance in the target variable was handled using Random Oversampling (ROS), ensuring a more balanced class distribution and reducing potential bias toward majority classes.

A Multilayer Perceptron (MLP) Classifier was employed as the prediction model. The model was trained using a supervised learning approach to classify maternal health risk levels based on clinical and physiological indicators. Performance evaluation was conducted using Accuracy, Precision, Recall, F1-Score, Specificity, and Classification Error.

RESULT

Data Quality Analysis

The initial dataset consisted of 1,205 maternal health records containing eleven predictor variables and one target variable (*Risk Level*). Missing values were identified in several attributes, including Systolic Blood Pressure, Diastolic Blood Pressure, Blood Sugar, BMI, Previous Complications, Preexisting Diabetes, Heart Rate, and Risk Level. To address this issue, linear interpolation was applied to numerical features, while mode imputation was used for the target variable.

Table 1. Missing Value Summary

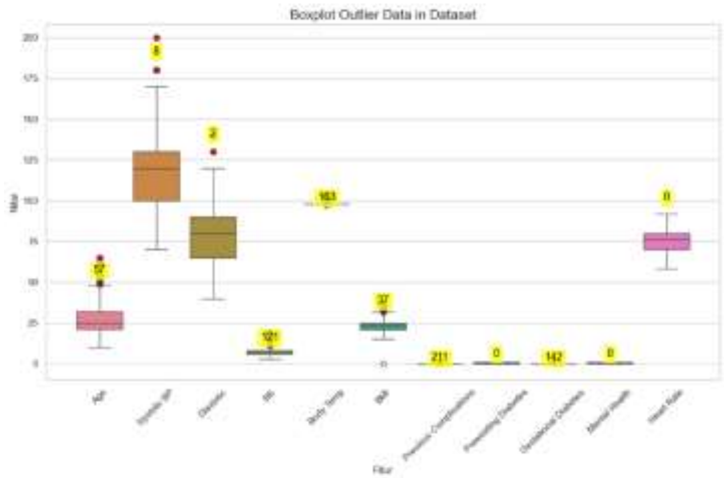
Missing Value Summary

No.	Variabel	Jumlah Missing Value (NaN)
1.	Age	0
2.	Systolic BP	5
3.	Diastolic	4
4.	BS	2
5.	Body Temp	0
6.	BMI	18
7.	Previous Complications	2
8.	Preexisting Diabetes	2
9.	Gestational Diabetes	0
10.	Mental Health	0
11.	Heart Rate	2
12.	Risk Level	18
Total		53

Duplicate analysis revealed the presence of redundant records. After duplicate removal, the dataset size decreased from 1,205 to 1,187 records, resulting in a cleaner dataset for subsequent analysis.

Outlier detection was performed using the Interquartile Range (IQR) method. Several variables exhibited outliers, particularly Previous Complications, Body Temperature, Gestational Diabetes, and Blood Sugar. Instead of removing these observations, the capping technique was applied to reduce the influence of extreme values while preserving the overall data distribution.

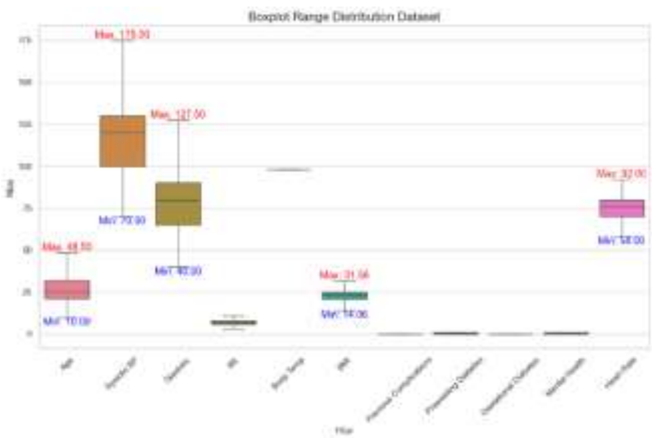
Figure 1. Boxplot of Outlier Detection Results



Feature Scaling

The dataset contained features with different numerical ranges, which could potentially affect model learning. Therefore, Robust Scaler was applied to normalize the data while maintaining resistance to extreme values. This preprocessing step helped ensure that all variables contributed proportionally during model training.

Figure 2. Feature Range Distribution Before Scaling

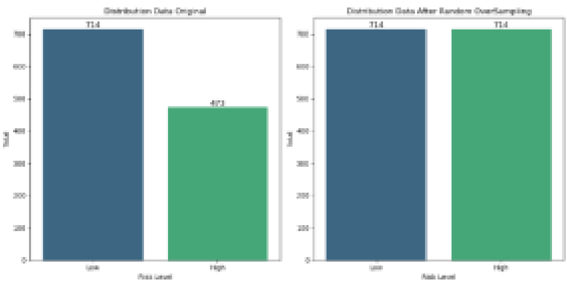


Class Balancing Results

The distribution of maternal health risk categories was found to be imbalanced. The original dataset contained 714 Low-Risk cases and 473 High-Risk cases. Such imbalance may negatively affect classification performance because machine learning algorithms tend to favor the majority class during training.

To address this issue, Random Oversampling (ROS) was applied to increase the representation of the minority class. The balancing process produced a more uniform class distribution, enabling the model to learn patterns from both risk categories more effectively.

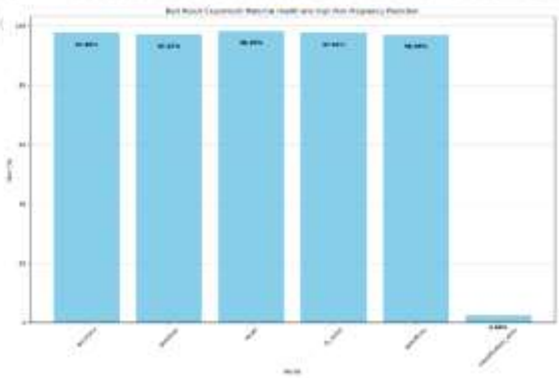
Figure 3. Distribution of Original Data and Data After Random Oversampling



Classification Performance

The performance of the proposed Multilayer Perceptron (MLP) Classifier was evaluated using Accuracy, Precision, Recall, F1-Score, Specificity, and Classification Error. The experimental results demonstrate that the proposed model achieved excellent classification performance.

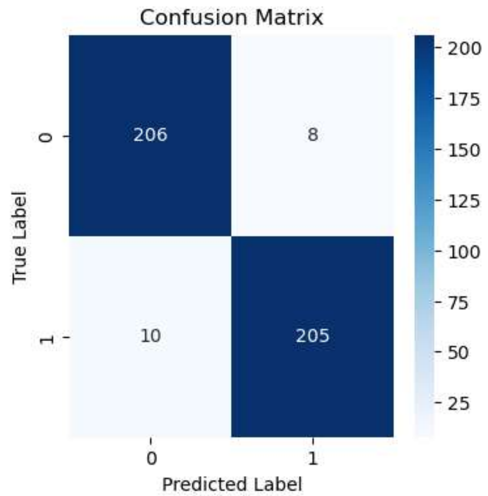
Figure 4. Best Results of Maternal Health Risk Prediction



The results indicate that the MLP Classifier was highly effective in predicting maternal health risk levels. An accuracy of 97.60% demonstrates that the majority of maternal health risk cases were classified correctly. The recall value of 98.20% indicates a strong ability to identify maternal risk conditions, while the precision value of 97.07% reflects a low false-positive rate. Furthermore, the F1-score of 97.62% confirms a balanced trade-off between precision and recall.

In addition, the specificity value of 96.99% shows that the model effectively distinguished low-risk from high-risk cases. The low classification error of 2.40% indicates the robustness and reliability of the proposed approach. These findings suggest that combining comprehensive preprocessing techniques with the MLP Classifier can significantly improve maternal health risk prediction and support early detection in healthcare settings.

Figure 5. Confusion Matrix of the MLP Classifier



The confusion matrix further confirms the effectiveness of the proposed model. Among 214 actual Low-Risk cases, 206 were correctly classified, while only 8 were incorrectly predicted as High Risk. Similarly, among 215 actual High-Risk cases, 205 were correctly identified, whereas 10 cases were misclassified as Low Risk. These findings demonstrate that the MLP Classifier achieved balanced performance across both classes and was able to accurately distinguish between low-risk and high-risk maternal conditions. The relatively small number of misclassified cases supports the high accuracy, recall, precision, and specificity obtained during model evaluation.

DISCUSSION

The findings indicate that preeclampsia was significantly associated with stunting incidence. Preeclampsia can impair placental perfusion, reducing the transfer of oxygen and nutrients to the fetus, which may lead to intrauterine growth restriction and subsequently increase the risk of stunting during childhood.

In contrast, maternal anemia was not significantly associated with stunting in this study. This finding differs from several previous studies that reported anemia during pregnancy as a risk factor for impaired fetal growth. The discrepancy may be related to differences in sample size, severity of anemia, nutritional interventions, and socioeconomic factors among study populations.Similarly, hemorrhage during pregnancy did not show a significant association with stunting. This result suggests that hemorrhage alone may not directly influence child growth outcomes, particularly when adequate medical management is provided.These findings contribute to the understanding that hypertensive disorders during pregnancy, especially preeclampsia, remain important determinants of child nutritional status. Therefore, strengthening antenatal care services for early detection and management of preeclampsia may help reduce the burden of stunting.

CONCLUSIONS

This study successfully developed an Artificial Intelligence (AI)-based maternal health risk prediction model using the Multilayer Perceptron (MLP) Classifier. The model was developed using maternal clinical and physiological indicators, including age, blood pressure, blood sugar level, body mass index, previous complications, diabetes history, mental health status, and heart rate. To enhance data quality and model reliability, several preprocessing techniques were applied, including missing value imputation, duplicate removal, outlier treatment using the capping method, Robust Scaler normalization, and Random Oversampling (ROS) for class balancing.

Future studies should utilize larger and more diverse datasets collected from multiple healthcare institutions to improve model generalizability and reduce potential population bias. Additionally, comparative evaluations involving other machine learning and deep learning algorithms may provide further insights into optimizing maternal health risk prediction performance. The integration of explainable AI techniques is also recommended to enhance model transparency and support healthcare professionals in interpreting prediction results.

ACKNOWLEDGEMENT

The author would like to express sincere gratitude to the Department of Informatics, Faculty of Information Technology and Data Science, Universitas Sebelas Maret, for providing academic support during this research. The author also thanks all parties who contributed to the completion of this study.

CONFLICT OF INTEREST AND FUNDING DISCLOSURE

The author declares that there is no conflict of interest regarding the publication of this article. This research received no specific grant from any funding agency in the public, commercial, or not for profit sectors.

AUTHOR CONTRIBUTIONS

Dien Kartika Lutfi: Conceptualization, methodology, data curation, formal analysis, investigation, validation, visualization, writing original draft, and writing review and editing.

REFERENCES

Al Mashrafi, S. S., Tafakori, L., & Abdollahian, M. (2024). Predicting maternal risk level using machine learning models. *BMC Pregnancy and Childbirth*, *24*(1). https://doi.org/10.1186/s12884-024-07030-9

Chayan, A. R. (2024). *Maternal Health and High-Risk Pregnancy Dataset*. https://doi.org/10.17632/8k9pvpmykk.1

Gao, J., Yao, Y., Xue, J., Chen, R., Yang, X. Y., Xu, J., & Cheng, W. (2025). Methodological conduct and risk of bias in studies on prenatal birthweight prediction models using machine learning techniques: a systematic review. *BMC Pregnancy and Childbirth*, *25*(1). https://doi.org/10.1186/s12884-025-07727-5

Islam, S., Shahriyar, R., Agarwala, A., Zaman, M., Ahamed, S., Rahman, R., Chowdhury, M. H., Sarker, F., & Mamun, K. A. (2025). Artificial intelligence-based risk assessment tools for sexual, reproductive and mental health: a systematic review. *BMC Medical Informatics and Decision Making*, *25*(1). https://doi.org/10.1186/s12911-025-02864-5

Khadidos, A. O., Saleem, F., Selvarajan, S., Ullah, Z., & Khadidos, A. O. (2024). Ensemble machine learning framework for predicting maternal health risk during pregnancy. *Scientific Reports*, *14*(1), 1–21. https://doi.org/10.1038/s41598-024-71934-x

Li, T., Xu, M., Wang, Y., Wang, Y., Tang, H., Duan, H., Zhao, G., Zheng, M., & Hu, Y. (2024). Prediction model of preeclampsia using machine learning based methods: a population based cohort study in China. *Frontiers in Endocrinology*, *15*(June), 1–14. https://doi.org/10.3389/fendo.2024.1345573

Malde, A., Prabhu, V. G., Banga, D., Hsieh, M., Renduchintala, C., & Pirrallo, R. (2025). A Machine Learning Approach for Predicting Maternal Health Risks in Lower-Middle-Income Countries Using Sparse Data and Vital Signs. *Future Internet*, *17*(5). https://doi.org/10.3390/fi17050190

Malik, V., Agrawal, N., Prasad, S., Talwar, S., Khatuja, R., Jain, S., Sehgal, N. P., Malik, N., Khatuja, J., & Madan, N. (2024). Prediction of Preeclampsia Using Machine Learning: A Systematic Review. *Cureus*, *16*(12). https://doi.org/10.7759/cureus.76095

Mamun, M., Chowdhury, S. H., Faruq, M. O., Azad, M. M., Biswas, B. R., Hussain, M. I., & Hossain, M. M. (2025). Identification of Maternal Health Risk From Optimal Features Using Explainable Machine Learning. *Engineering Reports*, *7*(11). https://doi.org/10.1002/eng2.70491

Özcan, M., & Peker, S. (2025). Preeclampsia prediction via machine learning: a systematic literature review. *Health Systems*, *14*(3), 208–222. https://doi.org/10.1080/20476965.2024.2435845

Pavagada, K., & Vemuri, J. (2025). Prediction of Maternal Health Risk Factors Using Machine Learning Algorithms. *Procedia Computer Science*, *258*, 2713–2722. https://doi.org/10.1016/j.procs.2025.04.532

Tzimourta, K. D., Tsipouras, M. G., Angelidis, P., Tsalikakis, D. G., & Orovou, E. (2025). Maternal Health Risk Detection: Advancing Midwifery with Artificial Intelligence. *Healthcare (Switzerland)*, *13*(7). https://doi.org/10.3390/healthcare13070833

Vasudevan, L., Kibria, M. G., Kucirka, L. M., Shieh, K., Wei, M., Masoumi, S., Balasubramanian, S., Victor, A., Conklin, J. L., Gurcan, M. N., Stuebe, A. M., & Page, D. (2025). Machine Learning Models to Predict Risk of Maternal Morbidity and Mortality From Electronic Medical Record Data: Scoping Review. *Journal of Medical Internet Research*, *27*, 1–13. https://doi.org/10.2196/68225

WHO. (2025). *Maternal Mortality*.

Xinyu Pi, et al. (2021). Prediction of high-risk pregnancy based on machine learning algorithms. *International Eye Science*, *21*(9), 1644–1648. https://doi.org/https://doi.org/10.1038/s41598-025-00450-3 1