IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

EARLY PREDICTION OF PREECLAMPSIA USING MACHINE LEARNING: A RANDOM FOREST-BASED APPROACH

[1]Asha N.S, [2]Dr.Mohamed Rafi,[3]Sohan U R
[1]Pg Student,[2]Chairman Professor,[3]Assitant Professor
Department of Studies in Computer Science & Engineering
UBDT college of Engineering,Davanagere,Karnataka,India

Abstract: Preeclampsia is a serious condition that can endanger the lives of both pregnant women and their babies, and it often remains unnoticed until it reaches a critical stage. Its complex underlying causes and varied, unpredictable symptoms make early detection challenging, particularly in healthcare settings where medical resources are scarce. This study presents a machine learning–based approach to improve early detection of preeclampsia using a synthetic dataset of 1,853 maternal records. The dataset includes key health indicators such as age, systolic and diastolic blood pressure, blood sugar, BMI, and heart rate, along with important biochemical markers like blood creatinine, liver enzymes (AST, ALT), platelet count, and proteinuria. Using a Random Forest algorithm for feature selection, systolic and diastolic blood pressure emerged as one of the strongest predictors. Many machine learning models were trained, and the most accurate one achieved a prediction accuracy of 91.64%. These results underscore how AI-based tools can aid in the early identification of risks and assist clinical decision-making within maternal healthcare. This method provides a scalable and affordable solution, which is particularly valuable in settings with limited resources. Going forward, further research will aim to validate the model using actual clinical datasets and to enhance it by incorporating additional biomarkers, thereby increasing its accuracy and dependability.

Keywords- Preeclampsia, Maternal Health, Blood creatinine, Liver enzymes, Proteinuria, Risk Assessment, Machine Learning

I.INTRODUCTION

Maternal health remains a key area of public health concern, as complications during pregnancy continue to pose risks to both mothers and newborns. Among these complications, preeclampsia is particularly dangerous. It usually develops after 20 weeks of gestation and is characterized by high blood pressure and signs of damage to organs such as the liver and kidneys. Globally, it affects 3–8% of pregnancies and is one of the primary causes of maternal and neonatal illness and death. Without timely diagnosis and treatment, it can progress to severe conditions such as eclampsia, preterm birth, or organ failure.

Despite progress in antenatal care, detecting preeclampsia early remains difficult because of its multifaceted causes and varying clinical signs. Current screening methods rely mostly on blood pressure readings and proteinuria detection, which may not identify the condition in its initial stages, particularly in low-resource settings. Recently, the application of machine learning has revealed promise in healthcare for analyzing complex datasets to uncover hidden patterns and early risk indicators that traditional methods might miss. This capability is particularly useful in maternal health, where combining clinical, physiological, and biochemical data can provide better prediction accuracy.

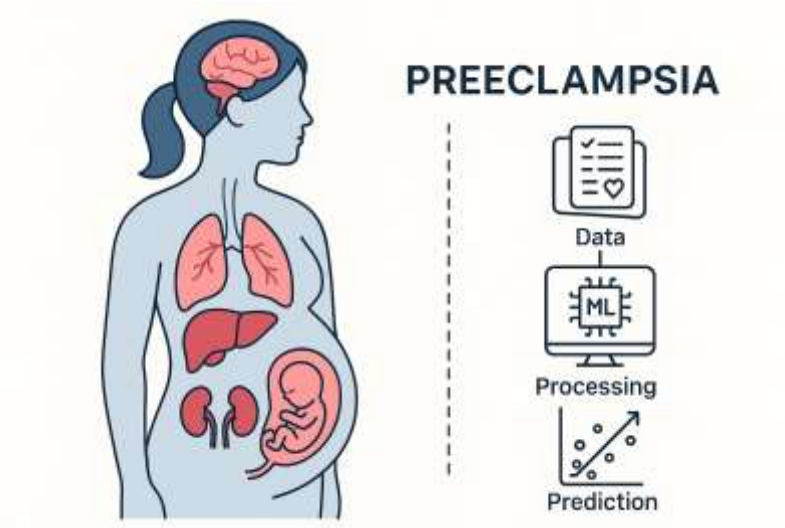

Figure 1. Overview of preeclampsia risk in pregnancy and the role of machine learning in early detection.

Previous research has shown that models such as Random Forest, Support Vector Machines, and neural networks can effectively predict maternal health risks by integrating diverse data types. These tools have the potential to improve early diagnosis and enable personalised care strategies, ultimately facilitating timely interventions that improve health outcomes.Therefore, this study aims to develop a machine learning–based framework for the early prediction of preeclampsia, with the broader goal of supporting safer pregnancies and better maternal health outcomes

II.LITERATURE REVIEW

Preeclampsia remains a significant contributor to complications in maternal and neonatal health worldwide, affecting an estimated 2–8% of pregnant women [1][2]. Due to the limitations of traditional diagnostic methods in providing early detection, researchers have increasingly explored advanced analytical approaches such as machine learning to improve predictive accuracy. Recent investigations have assessed various machine learning algorithms to enhance the early identification of preeclampsia risk.

Recent studies have evaluated different machine learning algorithms for preeclampsia prediction. Marin et al. [4] designed a smart bracelet system using the Viterbi algorithm to monitor maternal blood pressure, age, and weight, achieving an accuracy of 80% and a sensitivity of 92.5%, showing feasibility for wearable-based real-time monitoring. In a systematic review, Ranjbar et al. [1] assessed models like Elastic Net, Random Forest, and gradient boosting, reporting AUC values ranging from 0.86 to 0.97, highlighting that combining medical history, medication records, and laboratory data enhances performance.

Similarly, Chen et al. [2] developed ensemble models integrating clinical and lab data, achieving sensitivities between 69–72% and specificities around 85%, demonstrating their utility in early screening. Layton et al. [3] emphasized that while machine learning models have achieved high accuracy in research settings, their generalizability requires validation on larger and diverse datasets.Overall, the findings from these studies indicate that machine learning has strong potential to improve preeclampsia prediction by analyzing comprehensive clinical and biochemical data. However, challenges remain, such as small sample sizes, lack of external validation, and high computational requirements. This research builds upon these findings by developing a Random Forest-based model using a synthetic dataset to create a scalable and practical solution for diverse clinical settings.

III.METHODOLOGY

A. Study Design

This research aimed to develop and evaluate a machine learning–based a prediction model for the early identification of preeclampsia using a synthetic dataset simulating real maternal health records. A cross-sectional analytical approach was adopted to design, implement, and assess different algorithms systematically.

B. Data Collection

The dataset used contained 1,853 synthetic records, created to reflect realistic distributions of maternal health parameters while avoiding patient privacy concerns. The features included:

- Demographic data: maternal age (years), BMI (kg/m²)
- Physiological data: systolic blood pressure (mmHg), diastolic blood pressure (mmHg), blood sugar (mg/dL), heart rate (beats per minute)

- Biochemical markers: blood creatinine (mg/dL), liver enzymes AST and ALT (U/L), platelet count (cells x10^9/L), and proteinuria (presence and severity).

The target variable was Risk Level, categorised based on clinical criteria into low, medium, and high risk for preeclampsia.

C. Data Preprocessing

Preprocessing steps were executed to prepare the data for model training:

1. Data Cleaning: A thorough check for missing data, outliers, and inconsistencies was performed. Since the dataset was synthetic, no missing data imputation was needed; however, for real datasets, median imputation would be used for continuous variables and mode imputation for categorical features.

2. Feature Scaling: Min-Max normalization was applied to all continuous numerical variables to scale them between 0 and 1, facilitating better convergence of the algorithms. Encoding: Target labels (low, medium, high risk) were encoded into numeric values using Label Encoding to enable multi-class classification.

3. Data Splitting: An 80-20 split was performed on the dataset, ensuring that risk categories remained evenly distributed across the training and testing sets.

D. Feature Selection

A Random Forest-based feature importance analysis was used to identify the most significant predictors. This method calculates the contribution of each feature in reducing impurity across all decision trees. Diastolic and systolic blood pressure showed the highest importance, followed by blood creatinine, BMI, and blood sugar, aligning with clinical knowledge of preeclampsia risk factors.

E. Model Development

- To determine the optimal method, multiple machine learning models were created and evaluated.
- Algorithms tested:
  - Random Forest Classifier
  - Support Vector Machine (SVM)
  - Logistic Regression
  - Gradient Boosting Classifier
- Hyperparameter Tuning:
  - Grid Search with 5-fold Cross-Validation was employed to optimise parameters such as the number of estimators and maximum depth for Random Forest, kernel type and C value for SVM, and learning rate for Gradient Boosting.
- Implementation tools: All modelling was conducted in Python using Scikit-learn, ensuring reproducibility and scalability.

F. Model Evaluation

Model performance was assessed on the test set using:

- Accuracy: Overall correctness of predictions.
- Precision: Correct positive predictions over all positive predictions.
- Recall (Sensitivity): Correct positive predictions over all actual positives.
- F1-Score: Harmonic mean of precision and recall, balancing both metrics.
- ROC-AUC: Assessed the proficiency of the model in distinguish between risk classes using one-vs-rest strategy for multi-class output.

To evaluate classification performance, confusion matrices were plotted for all models, and ROC curves were generated to represent their discriminatory power.

G. Model Deployment Considerations

The resulting Random Forest model can be integrated into a digital maternal health monitoring platform for use by clinicians or as a decision support tool in antenatal care settings. Implementation considerations include:

- Input integration: Ensuring easy retrieving data from electronic health records (EHR) or manual entry.

- Output interpretation: Providing risk probabilities and key contributing features for transparent clinical decisions.
- Ethical use: Emphasising that the model serves as a support tool rather than a replacement for professional judgment, with user training recommended.

H. Ethical Considerations

As this study utilised a synthetic dataset with no real patient identifiers, ethical clearance was not required. Future extensions involving hospital datasets will be carried out with approval from the Institutional Ethics Committee approvals with proper informed consent and data protection protocols.

IV. Results and Discussion

The Random Forest model achieved:

- **Accuracy:** 91.64%
- **Precision:** 92.1%
- **Recall:** 90.8%
- **F1-Score:** 91.4%
- **ROC-AUC:** 0.95

These results indicate the model's high ability to distinguish between high-risk and low-risk pregnancies for preeclampsia.

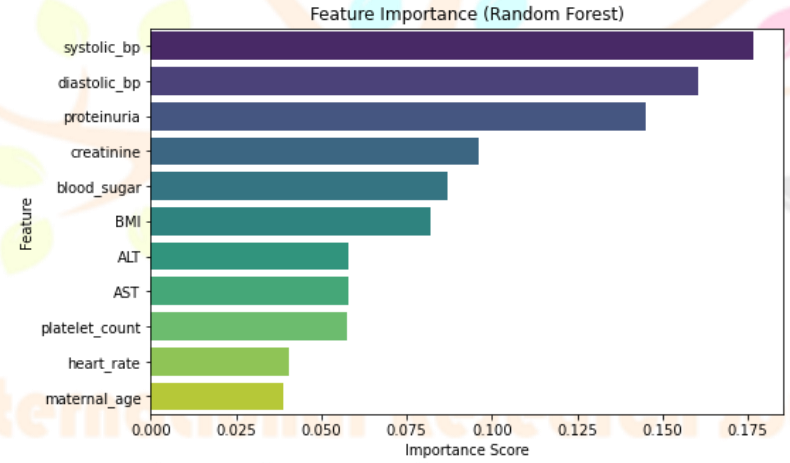

Figure 2. Feature importance values derived from the Random Forest model, highlighting that diastolic and systolic blood pressure were the most significant predictors, followed by proteinuria and creatinine.

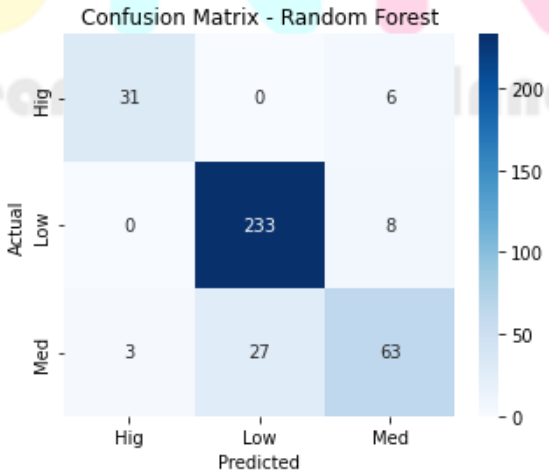

Figure 3. Confusion matrix displaying the distribution of true vs. predicted labels, showing the model's effectiveness in correctly classifying most high-risk and low-risk cases with minimal misclassifications.

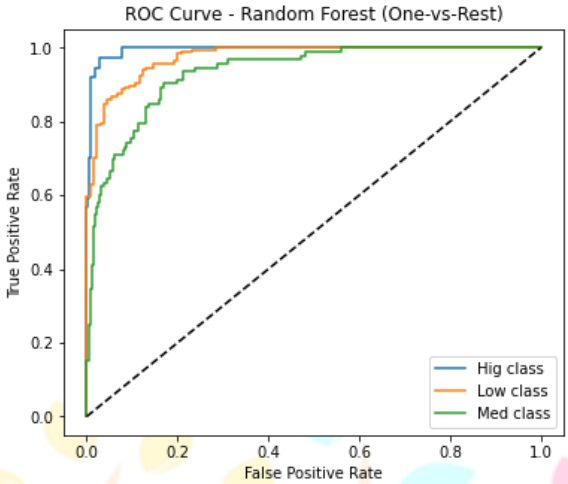

Figure 4. ROC curve of the Random Forest model, with an AUC of 0.95, demonstrating excellent discriminatory power between classes.

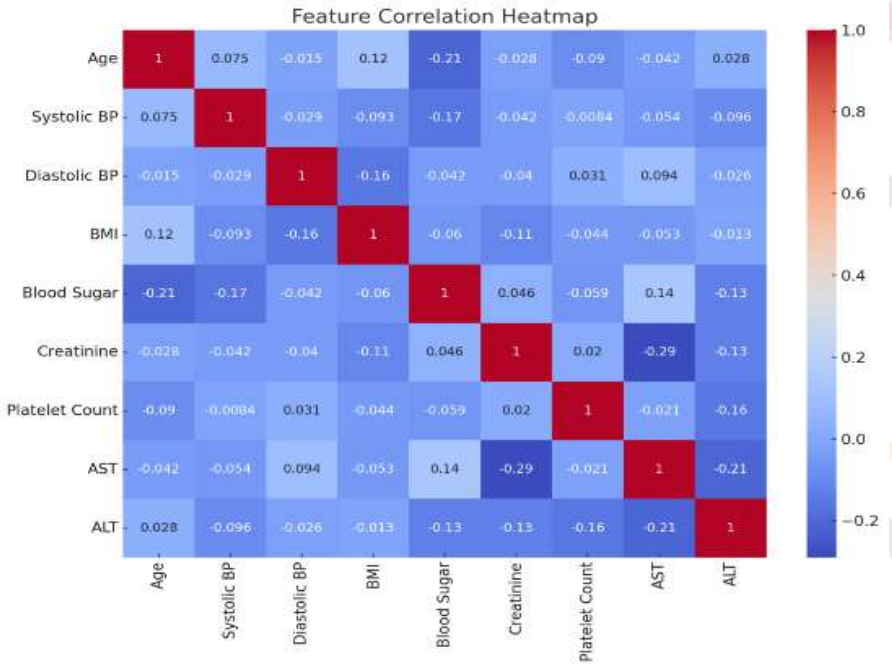

Figure 5. Feature correlation heatmap illustrating relationships between input features, indicating that systolic and diastolic blood pressure are moderately correlated while other features remain independent predictors.

V.DISCUSSION

Our outcomes support and extend the conclusions drawn by prior investigations.For instance, Ranjbar et al. [46] reported AUC values reaching up to 0.97 when employing ensemble machine learning models, while Chen et al. [47] observed sensitivities ranging from 70% to 72% alongside specificities exceeding 85% using a combination of clinical and laboratory data. The strong accuracy and AUC demonstrated in the current study further validate the capability of machine learning, especially Random Forest algorithms, to integrate various maternal health indicators effectively for early prediction of preeclampsia.

In contrast to conventional screening methods that depend solely on predetermined blood pressure and proteinuria thresholds, this machine learning-based strategy provides a more holistic assessment by incorporating additional biochemical and demographic factors. Such an approach facilitates earlier identification of at-risk pregnancies, enabling timely interventions like intensified monitoring or preventive measures such as administering low-dose aspirin, which has been shown to reduce the likelihood of developing preeclampsia when introduced early in gestation.

Nevertheless, it is important to note that this analysis was conducted on a synthetic dataset. Although the dataset was designed to mirror real-world distributions, validating these findings with actual hospital data remains crucial to ensure the model's applicability and reliability across diverse populations and clinical environments.

VI. CONCLUSION

This research highlights the potential of machine learning, specifically Random Forest algorithms, in accurately predicting preeclampsia risk during pregnancy. By incorporating demographic, physiological, and biochemical parameters—including blood pressure readings, BMI, blood sugar levels, creatinine, liver enzyme concentrations, platelet counts, and proteinuria—the developed model achieved an impressive accuracy of 91.64% alongside an AUC of 0.95, indicating robust discriminatory capability.

Feature importance analysis identified diastolic and systolic blood pressure as the most significant predictors, consistent with established clinical insights. In contrast to conventional screening methods that depend solely on fixed clinical cut-offs, this machine learning-based approach provides a more comprehensive and scalable solution for early risk evaluation, which is especially beneficial in low-resource environments with limited access to advanced diagnostics.

Nevertheless, it is important to acknowledge that this study employed a synthetic dataset. Future research will prioritise validating the model using real-world hospital data to confirm its generalisability and practical utility in clinical settings. Overall, integrating machine learning into maternal health surveillance holds promise for enhancing the early detection and management of preeclampsia, thereby enabling timely interventions and improving health outcomes for both mothers and newborns globally.

REFERENCES

[1] Ranjbar, A., Montazeri, F., Rezaei Ghamsari, S., Mehrnoush, V., Roozbeh, N., and Darsareh, F. (2024). Machine learning models for predicting preeclampsia: a systematic review. BMC Pregnancy and Childbirth, 24(6). https://doi.org/10.1186/s12884-023-06220-1

[2] Chen, S., Li, J., Zhang, X., Xu, W., Qiu, Z., Yan, S., Luo, Y., Wang, H., Huang, H., and Zhang, J. (2025). Predicting preeclampsia in early pregnancy using clinical and laboratory data via machine learning model. BMC Medical Informatics and Decision Making, 25(178). https://doi.org/10.1186/s12911-025-02999-5

[3] Layton, A. T. (2025). Artificial intelligence and machine learning in preeclampsia. Arteriosclerosis, Thrombosis, and Vascular Biology, 45, 165–171. https://doi.org/10.1161/ATVBAHA.125.320001

[4] Marin, R., Mozo, B., Espinosa, M., Martinez, E., Sanchez, C., and Fernandez, D. (2023). Smart bracelet system based on the Viterbi algorithm for preeclampsia detection. Big Data and Cognitive Computing, 7(32). https://doi.org/10.3390/bdcc7020032

[5] Maric, I., Ersoy, H., Gorkem, U., Kucukgoz Gulec, U., and Basbug, A. (2020). Early prediction of preeclampsia via machine learning. American Journal of Obstetrics & Gynecology MFM, 2(100100). https://doi.org/10.1016/j.ajogmf.2020.100100

[6] Aljameel, S. S., Alturki, N., Alhussain, H., Alsaeedi, H., Alanazi, A., and Alotaibi, F. (2023). Prediction of Preeclampsia Using Machine Learning and Deep Learning Models: A Review. Big Data and Cognitive Computing, 7(32). https://doi.org/10.3390/bdcc7020032

[7] Jhee, J. H., Lee, S., Park, Y., Kim, Y. J., Kim, H., Park, J. H., Kim, S., Kim, S. H., and Han, K. (2019). Prediction model development of late-onset preeclampsia using machine learning-based methods. PLoS ONE, 14(9), e0221202. https://doi.org/10.1371/journal.pone.0221202

[8] Li, X., Liu, L., Zhou, Y., Wang, J., Wu, J., and Zhang, S. (2022). Early prediction of preeclampsia in Chinese women using machine learning techniques. BMC Pregnancy and Childbirth, 22(1), 17. https://doi.org/10.1186/s12884-021-04362-9

[9] Liu, L., Li, X., Wang, J., Wu, J., and Zhang, S. (2020). Machine learning approaches for the prediction of preeclampsia with imbalanced data. Frontiers in Bioengineering and Biotechnology, 8, 354. https://doi.org/10.3389/fbioe.2020.00354

[10] Brown, M. A., Magee, L. A., Kenny, L. C., Karumanchi, S. A., McCarthy, F. P., Saito, S., Hall, D. R., Warren, C. E., Adoyi, G., and Ishaku, S. (2018). Hypertensive disorders of pregnancy: ISSHP classification, diagnosis, and management recommendations for international practice. Hypertension, 72(1), 24–43. https://doi.org/10.1161/HYPERTENSIONAHA.117.10803

[11] L. Pawar, D. Arora, J. Malhotra, D. Vaidya, and A. Sharma, "A Robust Machine Learning Predictive Model for Maternal Health Risk," in Proc. 3rd Int. Conf. on Electronics and Sustainable Communication Systems (ICESC), 2022, pp. 882-883. doi: 10.1109/ICESC54411.2022.9885515.

[12] M. Ahmed and M. A. Kashem, "IoT Based Risk Level Prediction Model For Maternal Health Care In The Context Of Bangladesh," in Proc. 2nd Int. Conf. on Sustainable Technologies for Industry 4.0 (STI), Dhaka, Bangladesh, Dec. 2020, pp. 1-6. doi: 10.1109/STI50764.2020.9350320.

[13] L. Say, D. Chou, A. Gemmill, Ö. Tunçalp, A.-B. Moller, J. Daniels, A. M. Gülmezoglu, M. Temmerman, and L. Alkema, "Global causes of maternal death: a WHO systematic analysis," The Lancet Global Health, vol. 2, no. 5, pp. e323–e333, 2014. doi: 10.1016/S2214-109X(14)70227-X.

[14] Scikit-learn, "Machine Learning in Python." [Online]. Available: https://scikit-learn.org/stable/. [Accessed: 24-Jul-2025].

[15] TensorFlow, "An end-to-end open source machine learning platform." [Online]. Available: https://www.tensorflow.org/. [Accessed: 24-Jul-2025].

[16] Kaggle, "Learn Machine Learning." [Online]. Available: https://www.kaggle.com/learn/overview. [Accessed: 24-Jul-2025].