THE SCIENTIFIC TEMPER (2026) Vol. 17 (4): 6117-6130
E-ISSN: 2231-6396, ISSN: 0976-8653
Doi: 10.58414/SCIENTIFICTEMPER.2026.17.4.22
https://scientifictemper.com/
RESEARCH ARTICLE

Early Detection of Preeclampsia and Gestational Hypertension Using Machine Learning Techniques

R. Gomathi[1], Dr. K. Menaka[2*]

Abstract

Gestational hypertension is a maternal entanglement, manifested through increased blood pressure which may occur later than week 20 of gestation period and may result in severe snags viz. preeclampsia, preterm neonatal delivery and injury of obstetric organs. Early and precise prediction is important because early tracking and cure would protect the pregnant women and the fetus from great number of dangers. Computational classification models are crucial for forecasting hypertensive disorders in pregnant women with analysis of large volumes of clinical, demographic and physiological data. Among the different predictive characteristics, Systolic Blood Pressure (SBP) and Diastolic Blood Pressure (DBP) are of particular concern, since the aberrant changes of SBP and DBP are considered to be directly related to the presence of gestational hypertension and preeclampsia. The classification models that have been employed in this study in the analysis are the Mutual Information based classification, Recursive feature elimination (RFE), Lasso Regression based classification and Tree based classification. Moreover, Hybrid Feature Selection strategy has been suggested which is a mixture of the Filter Method (Mutual Information) and an Embedded Tree-Based Model, which enhances the predictive accuracy. All the models were measured by the important characteristics like F1-score, precision, accuracy, specificity, and sensitivity. The findings show that the suggested hybrid method was the most successful and recorded the best accuracy of 98% in all the datasets and an acceptable result in the training and testing data. The paper represents various machine learning classification algorithms to forecast gestational hypertension and preeclampsia. These are optimized models that can help in diagnosing the issues at an early stage and also improve the results of the maternal healthcare by taking into consideration the variables of SBP, DBP and other clinically significant variables.

Keywords: Machine Learning, Gestational Hypertension, SBP, DBP, Preeclampsia, Classification.

Introduction

Gestational hypertension is a common hypertensive pregnancy disorder that is usually diagnosed after the 20th week in pregnancy and is characterized by a high level of blood pressure. The patient might develop severe complications, including preeclampsia, preterm birth, placental insufficiency, and damage of maternal organs. These possible negative consequences indicate the need to pay special attention to the early diagnosis, lifelong monitoring, and proper prediction of blood pressure disorders in women in pregnancy.

The health of the mother is essential in ascertaining the health of women in the processes of pregnancy, birth and during the aftermath. These complications during pregnancy can be influenced by multiple physiological and clinical aspects like the age of the mother, the presence of hematological disorders, and the irregularities of the heart rate, as well as other health indicators. These maternal risk factors should be identified early in order to avoid negative maternal and fetal outcomes. Maternal health analysis and early risk identification during pregnancy is thus essential in decreasing the chances of complications during the prenatal, perinatal and postnatal stages. In the proposed study, it is proposed to use a new method of automatic prediction of maternal health risks related to pregnancy (Ali Raza et al.,2022)

[1]Research Scholar, PG & Research Department of Computer Science, Urumu Dhanalakshmi College, (Affiliated to Bharathidasan University), Tiruchirappalli, Tamil Nadu, India

[2]Head & Assistant Professor, PG & Research Department of Computer Science, Urumu Dhanalakshmi College, (Affiliated to Bharathidasan University), Tiruchirappalli, Tamil Nadu, India

***Corresponding Author:** Dr. K. Menaka, Head & Assistant Professor, PG & Research Department of Computer Science, Urumu Dhanalakshmi College, (Affiliated to Bharathidasan University), Tiruchirappalli, Tamil Nadu, India, E-Mail: k.menaka@udc.ac.in

How to cite this article: Gomathi, R., Menaka, K. (2026). Early Detection of Preeclampsia and Gestational Hypertension Using Machine Learning Techniques. The Scientific Temper, **17**(4):6117-6130.

Doi: 10.58414/SCIENTIFICTEMPER.2026.17.4.22

Source of support: Nil

Conflict of interest: None.

Received: 29/03/2026 **Accepted:** 17/04/2026 **Published:** 25/04/2026

Preeclampsia is one of the most common maternal long and deadly illnesses in all parts of the globe with early diagnosis being essential to minimize maternal and fetal related tragedies. This research focuses on developing a predictive model using Machine Learning to assess the probability of Preeclampsia from maternal factors such as body mass index, mean arterial pressure, and prior clinical history of hypertension or diabetes mellitus. Several popular classification algorithms were considered to determine the predictive performance. The findings have shown the potential promise of robust and understandable predictive systems to improve assessment of preeclampsia in the basic medical care environment (Dwirani Ameliaet al.,2025). Risky pregnancies are a serious issue that presents challenges in maternal healthcare, and in most cases, proper prediction of the risks is needed to avoid the negative consequences. Machine learning (ML) models provide a potentially effective way of forecasting such risks by providing the opportunity to intervene timely and inform the clinical decision-making process. The pregnancy is a crucial period of life of any woman, and in this period the maternal nutrition is one of the most important factors protecting the well-being of the mother and the unborn child. Proper nutrition throughout pregnancy does not only benefit the mother but also affects fetal and biological development, birth weight, and developmental milestones and mortality into childhood and adulthood. The proposed outcomes of this research are expected to positively impact maternal and child health by providing healthcare professionals with data collection and analysis tools that would allow them to provide more personalized care during pregnancy (Atik Shahariyar Hasan et al.,2024).

The principal objective of this study is to evaluate the predictive performance of a novel high-risk factor scoring model utilizing machine learning algorithms for the early prediction of preeclampsia. The systolic and diastolic blood pressures are considered most important target variables that are used to forecast preeclampsia. This study will include the discussion of the likelihood that the developed high-risk score coupled with the machine learning models can effectively detect the preeclampsia at the initial stage of pregnancy and the high-risk antenatal cases with the help of the massive sample size. Several ML algorithms were used in order to develop and examine the predictive models. The efficacy of these models in determining high-risk women of preeclampsia was critically analyzed by conventional performance measures (Seeta Devi et al.,2024).

The adoption of ML approaches to predict and classify the data can be successfully done using the enormous volume of information on patients in the healthcare setting. Through effective medical data processing and analysis of large volumes of data, some diseases can be mitigated and even avoided by identifying the risk factors in good time. These types of data-driven ML can assist medical practitioners predict diseases more precisely and quickly and assist in early disease detection, informed more accurate clinical decisions and improved patient outcomes (P. Vinnarasi, Dr.K. Menaka.,2025).

The primary focus of this research was to determine the blood-related features of the patients and investigate the relationship between the severe preeclampsia and the value of blood index, which supports the diagnosis and treatment of severe preeclampsia at its early stage. The indicators do not regularly feature in regular check-ups of maternal health, but the early evaluation can offer useful clinical information. Health officials in China advise pregnant women to develop ante-natal care in the first trimester and most women report to hospitals very early to undergo detailed physical examinations or because of pregnancy complications like nausea and vomiting. The necessities of early detection of abnormalities in these blood indices could be the foundation of effective interventions that could be timely and thus prevent or lessen the severity of preeclampsia (Xinyuan Zhanga et al.,2022).

The study demonstrated that abnormal changes in SBP and DBP levels have strong and direct connections with the onset of gestational hypertension and preeclampsia, thus making these parameters the paramount risk factors. The main aim of the study, is to compare and evaluate the efficacy of different techniques in predicting gestational hypertension with the use of systolic blood pressure (SBP) and diastolic blood pressure (DBP) and other pertinent clinical parameters. This paper has a practical implication in maternal care, given that the hypothetical simplified machine learning model can be used to detect gestational hypertension early on. Additionally, it is possible to conduct prediction and classification with ML methods by using the combination of big patient data in the healthcare sector. The processing of large medical databases becomes efficient, which allows disease prediction faster and more accurately, reducing prevalence of diseases by helping to diagnose them in time and make better clinical decisions.

Literature Review

This paper examines the various literatures available on the prediction of gestational hypertension with numerous computing techniques, the study that employs ML to envisage several types of diseases and explain the application of the available techniques in the clinical sectors. There are several classification models that have been widely studied in the literature to enhance predictive power with a few of them being the Mutual Information-based classification, Recursive Feature Elimination (RFE) classification, Lasso regression-based classification and tree-based classification model. To enhance further the sensitivity of prediction, mechanisms of hybrid feature selection that are dependent on combinations of filter and embedded methods have been proposed. More precisely, the synergistic combination of Mutual Information with embedded tree-based models shows better efficiency in terms of better feature relevance as well as feature interactions, thus enabling more robust and highly accurate predictive models.

Jonathan.S et al.,2024 aimed at developing and internally validating machine learning-based prediction of Obstetric hypertension during prenatal care initiation. The main outcome in that study was the occurrence of HDP, Random Forest algorithms were widely used to develop predictive models because of their strength and the possibility to work with complex clinical information. RFE was used to produce compact models of individual result by removing the least relevant predictive features to make it more efficient and interpretable. The results indicated that, when using low-risk nulliparous pregnant populations, HDP at the start of prenatal care was correctly predicted using ML-based prediction models. This early prediction can be closely monitored and prophylactic measures may be implemented in time so as to better the outcome of the mother and fetus.

Martínez-Velasco et al.,2018 suggested a type of machine learning based on the classification model optimization applied towards prediction of preeclampsia/eclampsia (PE) syndrome. To identify the most relevant patient attributes related to the disease, the model was improved by the means of selecting the features efficiently and optimizing the thresholds. In order to better understand the PE syndrome, the authors implemented a number of machine learning methods various ML approaches which were successful in medical diagnosis and disease prevention. The research established a strong system of classifications that had the potential to define the probability of a person developing preeclampsia and also established the importance of duration of completed pregnancy as one of its predictors.

Desy Nuryunarsih et al., 2025 explored some ML methods based on the tree approach to forecast the blood pressure changes associated with hypertensive patients with diverse comorbidity conditions. Hypertension is a multidimensional problem that poses a major burden on the cardiovascular disease and mortality in the world. This paper compared three algorithms, including Decision Tree (DT), Random Forest (RF), and eXtreme Gradient Boosting (XGBoost), for the prediction of SBP and DBP. The interpretable pathways were produced through the use of decision tree plots. These findings supported the efficacy of tree-based ML models at individualized treatment response prediction to assist clinical decision-making.

Seyedeh Somayyeh Mousav et al.,2024 created a ML model to estimate the likelihood of hypertensive disorders of pregnancy (HDP) and their predetermination in the future with the help of cost-effective methods. The model applied the characteristics of blood pressure (BP) measurements, body mass index (BMI) values in the first and second trimester, and mother's demographic details, which allowed to take risks early and introduce preventive actions in time. The model was successful in capturing the important patterns of features needed to predict HDP and can be utilized through the simple technologies of BP monitors and weight scales, which makes it an affordable option in a low-resource context. This method can inform interventions and enhance care regarding adverse pregnancy outcomes because it allows detection early in life.

Yanto, Gusrino et al.,2025 carried out the study on the risk of preeclampsia prediction among prenatal females in the community health center Anak Air in Padang City. The experiment was based on a random forest (RF) model and other methods therein explained the relevance of the parameters. The review established the important risk factors of pregnancy such as age, blood, temperature, diastolic blood pressure (BP), blood sugar (BS), heart rate. The research proves that the Random Forest models, along with the explainable AI methods, can be used to facilitate the early diagnosis of preeclampsia and improve the process of decision-making in maternal care.

Zhang Yang et al.,2025 constructed and tested Prognostic models for delayed manifestation preeclampsia based on general information of the patient, maternal risk factors, and laboratory data that were measured in the initial gestation. The study encompassed pregnancies, with 110 cases diagnosed as late-onset preeclampsia. The findings illustrated that all models performed better when laboratory indicators were added rather than only when risk factors were employed. Specifically, some models outperformed than logistic regression significantly in the identification of Preeclampsia, which implies the superiority of machine learning methods in predicting maternal complications earlier.

Tahir Muhlis et al.,2018 aimed at foretelling the danger of preeclampsia, within the framework of hypertension and proteinuria following a gestation of more than 20 weeks. The Particle Swarm Optimization (PSO) was used for improving the model's performances with high precision and less computation time. Deep Learning model also offered an end-to-end methodology, where only input and output data is needed without a massive amount of feature engineering, and it proved to be effective in working with complicated data. The results imply that both neural network and deep learning to predict the risk of preeclampsia and Deep Learning has practical benefits with respect to automation and throughput.

Antony Sekar et al., 2025 suggested a blended predictive model with the combination of popular algorithms to improve the performance of the discrete outcome prediction that takes into consideration the linear and the nonlinear association in the multifaceted datasets. Experiments were done in a scalable and transparent

Google Colab environment in Python. The findings proved that the hybrid model was more accurate, precise, recalls more, F1-score and AUC-ROC than single models. The research offers a set of facts suggesting that hybrid machine learning solutions were capable of providing solutions that are strong, interpretable, and computationally inexpensive, and the model created has potential to become an efficient tool in preeclampsia diagnosis.

Shyu. I et al., 2025 created a predictive model of ML to evaluate the menace of preeclampsia concerning clinical data that may be collected regularly. Multiple ML algorithms were trained on a 70: 30 train : validation split. The research used the Synthetic Minority Oversampling Technique (SMOTE) to overcome the issue of class imbalance. XGBoost was the most cost-effective and accurate among the models to predict preeclampsia risk, allowing the assessment of elements in real-time and the use of early intervention. The authors have observed that the further work should be directed to large datasets and their incorporation into clinical practice.

Krishnarjun Bora et al., 2025 inspected the relevance of ML tools in predicting and detecting hypertension, which is a disease affecting a significant percentage of the world population. The research compared several ML classifiers and discovered that, the Random Forest Classifier (RFC) worked surpasses other classifiers in prediction of hypertension and predicting risks. Analysis of feature importance indicated that an individual had the strongest predictive power of hypertension with the use of waist circumference and BMI, whereas baseline SBP, DBP and BMI were the most significant factors in 2-year hypertension risk prediction.

Dewangan, Sonal et al.,2025 emphasized preeclampsia as one of the most critical complications of pregnancy, characterized by hypertension, lower limb edema, proteinuria, and thrombocytopenia. This multisystem disorder has the potential to cause severe maternal and fetal complications. Their study identified renal disease, chronic hypertension, a family history of hypertension or diabetes, and blood pressure ≥120/80 mmHg in early pregnancy as the most significant predictors of preeclampsia. Among the evaluated models, RF algorithm demonstrated the chief effectiveness in screening and predicting preeclampsia, thereby enabling obstetricians to identify high-risk pregnancies at an early stage and minimize adverse outcomes.

Rachel Bennett et al., 2022 indicated preeclampsia (PE) a s one of the hypertensive complications, which is experienced every year amid 8-10 percent of pregnancies in the United States. Although PE lacks a definite treatment, the use of aspirin can be adopted to reduce the severity of complications among risky individuals. The research also found out that PE is predominant among the racial minority groups and causes undue morbidity and death. Their findings were used to demonstrate that in the case of minorities, predictive performance would be fluctuating through the use of their experimental findings. These findings demonstrated the developed evidence of the predictive ability of clinical databases in risk assessment of preeclampsia among minorities.

Jagruti Meshram et al., 2023 measured the efficacy of Gestosis scale, a proven instrument to assess women at complication of developing pregnancy associated hypertension with the help of machine learning methods. There was the prospective observational study, in which 70 pregnant women were put on the observational study, The clinical features of gestosis were classified as mild, moderate and severe. Adaptive Boosting (AdaBoost) model had better predictive results with an accuracy of 97-99% in regression and classification model and a true-positive of 90%. The authors also realized that Gestosis score is a simple and non-invasive tool that can be effectively applied by the front-line health workers. Another aspect of the study that the researchers noted was the deployment of Data Mining and ML methods to early PIH prediction, which would make it possible to intervene in time, increase clinical decision-making, and positively influence maternal care results.

Changxiu Wang et al., 2025 developed and checked a model which helped to aid for the premature intervention of the preeclampsia (PE) problem during medical practices. Univariate and multivariate analyses were applied to find the independent menace of PE and a prognostic framework based on Logistic Regression was built. Its ability to distinguish outcomes was assessed using the area under the ROC curve (AUROC), while the agreement between predicted and observed results was examined through calibration plots. The nomogram generated, based on familiar and understandable clinical characteristics, proved to be highly predictive, and it also provides a beneficial insight on the establishment of specific preventative measures in clinical practice.

Alaa O. Khadidos et al., 2024 dwelled upon the examination of the Maternal Health Risk (MHR) factors pertaining to the pregnancy complications due to the presence of such conditions as hypertension, abnormal glucose levels, depression, and anxiety. This paper was meant to identify and estimate those factors based on real-life data. To attain it, the authors suggested a "A hybrid quad-ensemble model for maternal health risk stratification, incorporating multiple machine learning algorithms and four ensemble methods to boost predictive capability". The models are learnt in a group learning state to enhance the power of classification. The experiment findings showed that the high-risk (HR) group was the most predictive with the highest evaluation measures and the appropriate rate of prediction was 0.90.

Xinyu Pi et al.,2025 utilized ML Algorithms for the prediction of identifying high-risk pregnancies, aiming to develop an effective predictive model to support improved maternal health management. This developed an effective pregnancy risk prediction model by using the MLP method.

The results of the model evaluation depicted that it produced precise results for finding the risks of pregnancy, with the general predictive accuracy of the model being 82 percent. The findings underscore the potential of ML in the medical field, particularly for the timely and accurate identification of at-risk pregnancies, supporting clinical decision-making and offering useful insights for subsequent studies.

Amin.P et al.,2025 studied the issue of preeclampsia, hypertensive pregnancy that has very serious consequences on the well-being of the mother and her unborn baby and the researcher emphasizes that it should be detected early to be able to predict it and treat it accordingly. The significant predictors were selected with the help various feature selection methods. A number of classification models have been derived with an immense focus on ensemble learning models to enhance prediction. The proposed Hybrid Smart Selection Method (HSSM) outperformed other approaches in terms of the highest classification accuracy 98 percent and AUC-ROC of 99 percent with CFS and 99 percent with RFE and PCA. The findings demonstrate that, the feature selection and generalization of models is a viable and holistic framework which will contribute to timely and informed clinical decision in the management of preeclampsia.

The article by Paula L. Hedley et al., 2023 is a systematic review that tries to list the existing applications of AI/ML methods in screening of PE in the early pregnancy stage. The review critically evaluated the articles that employ AI/ML-based risk prediction models and was supposed to support the progress of medically significant algorithms that can be applied to issue timely interventions and support the development of new treatment plans.

Harizahayu et al., 2024 designed a model for PE menace diagnosis based on biomedical datasets. The data set consisted of demographic factors, blood_pressure, weight, maternal_age, history of pre-eclampsia, BMI, parity and inheritance and environmental factors. The outcome variable was preeclampsia risk, either a binary variable or a risk score. The accuracy of the random forest model was about 65.22, but the Kappa was very low, which means that the model did not separate the classes well, which meant that further refinement of the features or hyper parameter optimization or some other modeling method could be used to rise the predictive score of the model.

Materials And Methods

The methodology of the research is aimed at determining the major maternal risks factors as well as implementing machine learning-based predictive models that can be used to detect preeclampsia early and determine the outcome of the disease. This research will contribute to better identifying GH and PE in pregnant females by increasing the accuracy of the classification process with the help of efficient feature reduction methods to be used on the data. The system design phase has a number of major steps involved, such as data gathering & preprocessing and the application of suitable classification algorithms, and extensive model training and testing so as to secure strong and dependable predictive performance.

Data Collection Process

The current study has used a dataset, which includes 1,630 records of pregnant women gathered in various healthcare facilities and made publicly available in Kaggle. There are coded levels of maternal risk in the dataset, which enables a precise analysis of the correlation of systolic blood pressure (SBP), diastolic blood pressure (DBP), and maternal risk status. The data set consists of 60 variables, including such demographic characteristics as client identification and maternity age and important clinical measurements, especially SBP and DBP. It, further, includes a broad spectrum of maternal health measures that comprise the status of pregnancy, medical history, reported symptoms, and clinical interventions thus offering a holistic base on which the risk factors and health outcomes of preeclampsia among mothers are analyzed.

Table 1 presents a comprehensive examination of the characteristics of SBP & DBP datasets used in this study. The data is a varied compilation of demographic, clinical, and blood pressure-related variables that are applicable in studying gestational hypertension and preeclampsia. The demographic items are patient identification and maternal age that give the needed background information of every record. The variables of the obstetric history represent the characteristics related to pregnancy and maternal reproductive history. Factors related to the antenatal care would show how medical services are available and the quality of care provided during pregnancy. Some clinical variables are maternal medical history and pregnancy-related hypertensive disorders, such as chronic hypertension, diabetes mellitus, convulsions, headache, blurred vision, edema, abdominal pain, and bleeding. Additional information about laboratory and diagnostic pointers, including hemoglobin levels, glucose testing, ultrasound testing, and iron-folic acid supplementation provides further information about maternal health status and clinical intervention. Blood pressure-specialized features comprise an essential part of the data and consist of raw SBP and DBP variables, calculated means, categorical ones (e.g., SBP ≥140 mmHg and DBP ≥90 mmHg), and difference features.

The variables serve to detect the elevated blood pressure patterns and determine their relationship with gestational hypertension and preeclampsia. Gestational hypertension, preeclampsia and other related outcomes in the diagnostic variables show the presence and severity of the disease and the severe and superimposed cases. In general, the features enumerated in Table 1 facilitate the in-depth examination of maternal risk factors and blood pressure changes in pregnancy, which forms a strong basis

Table 1: SBP and DBP Variable Description of the Maternal Health Risk Dataset

Input Variable	*Attribute*	*Definition*
User-id	Arithmetic	Patient ID
PatientA	Arithmetic	Patient Age
Gestational Age	Arithmetic	Time duration of pregnancy in week
Prenatal visit	Categorical	Maternal Healthcare Access
Birth History	Arithmetic	A set of facts
Pregnancy Count	Arithmetic	Total number of deliveries past 20weeks
Mage	Categorical	Mother Age of group
Gravida Group	Categorical	Number of Pregnancy
GA-category	Arithmetic	Gestational age category
GA at 20 weeks	Arithmetic	Specific time point in pregnancy
GA Category 2	Categorical	Second level or alternative classification scheme
ANC Category	Categorical	Number or pattern of antenatal care
History of hypertension	Arithmetic	High blood pressure that develops gradually over time
Diabetes History	Arithmetic	Metabolic disorder characterized by elevated levels
Seizure	Arithmetic	Neurological event resulting from abnormal
Head discomfort	Arithmetic	A pain located in the head
Visual Disturbance	Arithmetic	Eye issue is often referred to by these alternative names
Bump	Arithmetic	Swelling on the body from a blow
Gut ache	Arithmetic	Symptom Characterized by abdominal pain
Anger evaluation	Categorical	Systematic assessment of a person's anger levels
Guidance	Categorical	The process of providing advice
Preeclampsia warning advice	Arithmetic	Early Detection of preeclampsia
Blood Hemoglobin	Categorical	Complex iron containing protein in erythrocytes
Gestational Diabetes	Categorical	That develops during pregnancy
Blood sample	Arithmetic	A small amount of blood
Blood Glucose	Categorical	Concentration of glucose present in the bloodstream
Ultersound scan	Categorical	High frequency sound waves
Iron folic acid tablet	Categorical	Medicinal supplement containing iron and folic acid
Calcium	Categorical	Mineral and Electrolyte
Magnesium Sulfate IM	Categorical	Maternal emergency care in hypertensive disorders of pregnancy.
Magnesium Sulfate IV	Categorical	Critical intervention in maternal emergency care
Valium	Categorical	Medication belonging to the benzodiazepine class
Antihypertensive	Categorical	Lower high blood pressure
Systolic Blood Pressure	Arithmetic	Classification of systolic blood pressure
Average SBP	Arithmetic	Mean value of all valid systolic blood pressure
SBP Valid Category	Arithmetic	Monitor blood pressure patterns during pregnancy
Diastolic Blood Pressure	Arithmetic	Identify women at risk for gestational hypertension or preeclampsia
Average DBP	Arithmetic	Mean value of all valid diastolic blood pressure
DBP Valid Category	Arithmetic	Classification of valid diastolic blood pressure
Valid Diagnosis	Categorical	Clinically confirmed and verified medical diagnosis
Initial Diagnosis	Categorical	During the first gestational week
First Visit BP Category	Categorical	Monitor early risk of gestational hypertension
Initial BP Unit	Categorical	Measurement unit used for blood pressure at the first visit
Valid Hypertension	Categorical	Clinically confirmed diagnosis of high blood pressure
Valid Preeclampsia	Categorical	Clinically confirmed diagnosis of preeclampsia

Cont...

Initial Hypertension	Arithmetic	Hypertension status recorded at the patient's
Preeclampsia Status at First Visit	Arithmetic	Preeclampsia status recorded at the patient's
Initial Hypertension BP	Arithmetic	Blood pressure measurement used to determine hypertension status
Initial Preeclampsia BP	Arithmetic	Blood pressure measurement used to assess preeclampsia status
Initial Hypertension BP Unit	Arithmetic	Unit of measurement used for recording blood pressure
First Visit Preeclampsia BP Unit	Arithmetic	Early detection and monitoring of preeclampsia risk
SBP Change	Arithmetic	Change or difference in systolic blood pressure
DBP Change	Arithmetic	Change or difference in diastolic blood pressure
Absolute SBP Change	Arithmetic	Absolute value of the change in systolic blood pressure
Absolute DBP Change	Arithmetic	Absolute value of the change in diastolic blood pressure
High DBP	Arithmetic	Diastolic blood pressure readings that exceed the normal threshold
Secondary Valid Diagnosis	Arithmetic	Clinically confirmed follow-up
Tertiary Valid Diagnosis	Arithmetic	Clinically confirmed diagnosis recorded during a later follow-up or evaluation
First Visit Diagnosis 2	Arithmetic	Diagnosis recorded during a follow-up evaluation
Magnesium Sulfate	Arithmetic	Medication used to prevent and treat seizures in pregnant women with severe preeclampsia

of feature selection, classification and predicting gestational hypertension and preeclampsia using machine learning.

Data Pre-Processing

In the case of data preprocessing, the loading of SBP and DBP data sets is done by loading Panda's library. Missing data that were captured by the character NaN have initially been filled with the value NaN that the missing data can be treated normally and the systematic imputations are simpler to make the values of the SBP and DBP numerical features that have missing values such as id,age,ga_wk,para,gravida,ga_20wk,hx_htn,hx_htn,hx_dm,seizure,headache,blur,edema,abd_pain,pe_counsel,htn_fu,pe_fu,high_dbp,dx2,dx3,dx_fu,mgso4,target and other blood pressure measurements were filled in with the average value of the concerned feature. In the case of the categorical features,such as anc visits, age cat,parity cat,gravida cat, ga cat,ga cat2,anc cat,danger assess,danger counsel,hb level,gdm status,glucose,usg,ifa,calcium,magsulf_im,magsulf_iv,diazepam,anti-htn,sbp fu cat,sbp valid cat,dbp fu cat,dbp cat,dx value,dx forward,dx fu bp,dx fwd bp u. The technique of label encoding has been applied to all nominal and ordinal variables to change them into quantitative variables since ML algorithms demand numerical inputs. After that, the processed dataset was segregated into training and testing portions according to the 80: 20 splits allowing the effective training of the models and evaluating their performance correctly.

Machine Learning (ML)

ML is a division of AI that enables the creation of schemes which may acquire using the available data. ML allows the systems to predict the future occurrences based on the previous experiences. ML creates models by training them on data to determine meaningful patterns and produce correct prediction. As such, the given work is devoted to developing and deploying an ML model that will be capable of predicting the levels of SBP & DBP. The procedure is done by utilizing ML algorithms to identify trends and relationships that exist in the data. In addition, the dataset is preprocessed to deal with the missing values and guarantee the data quality and consistency (Shahadat Uddin et al., 2026).

Supervised Learning

It is a ML paradigm in which the training records are labeled, i.e. the input variables and the target values are familiar. These algorithms are applied in the context of SBP and DBP prediction, where the clinical and demographic input variables are learnt to predict the known values of SBP and DBP.

Classification

Classification is a trained machine learning method that categorizes data instances within a set of already known data categories and it learns relationships between two or more data features. In the research, preeclampsia would be forecasted by the classification of pregnant women into a category, depending on the condition of gestational hypertension status. This study presents a relative assessment of four widely used ML classification algorithms in forecasting disease consequences, Thakur Rajneesh, et al., (2023).

Implemented Classification Models

Various classification models were applied in this paper by employing various feature selection methods. The approaches that were embraced included:

- Filter Method -Mutual Information (MI)-Based Classification.

- Wrapper Method Feature Elimination (RFE) based Classification by recursion.
- Embedded Approach - Tree Based Feature Selection and Classification.
- Hybrid Feature Selection Method Filter (Mutual Information) + Embedded (Tree-Based) Classification.

Filter Method Mutual Information -MI Classification

For measuring the statistical association of individual features and the response variable, mutual information is generally utilized. Attributes that have higher MI values give more information therefore they are picked to classify.

Wrapper Method Recursive Feature Elimination of Classification

Recursive Feature Elimination is an iterative method, that repeatedly removes less useful features based on how well the classifier performs. The outcome of the process is a subset of features that are optimal and improves the level of classification.

Embedded Approach – Tree Based Classifier

The nature of tree-based classifiers is such that they do feature selection when training models by the assignment of importance scores to features by the reduction of impurity or information gain, thus determining the most influential features.

Hybrid Feature Selection -Filter (Mutual Information) + Embedded (Tree-Based) Method

Under the hybrid strategy, firstly, Mutual Information is used to reduce the dimensions by selecting the most informative features. This is followed by splitting the feature subset where a tree-based embedded approach is used to further reduce it, which produces better classification and model robustness.

The approaches chosen to be applied to the dataset are the following:

Filter method (Mutual Information) + embedded method (Tree Based)

Results And Discussion

In this comparative analysis, Python was used to develop the analytical models in Jupiter Notebook within the Anaconda environment, facilitating efficient data exploration, visualization, and pattern identification. Multiple classification algorithms were evaluated to determine the most suitable model for the dataset. Machine Learning Repository provided by the health-related data was divided into training and testing set to develop and evaluate the model. The dataset was prepared by pre-processing, such as the task of dealing with missing values and the normalization of the data. Before the classification, feature selection methods were used in order to find the most important characteristics and improve the work of the models.

Algorithm: Selected feature method (Filter (Mutual Information) + embedded (Tree Based)):

Input:

- Dataset (df)

Process:

1. Load the dataset
2. Replace '?' with NaN
3. Handle missing values: numerical columns
4. Splitting up of data as training and testing sets
5. **Filter Method: Mutual Information**
6. **Compute Mutual Information**

   i) Calculate Mutual Information (MI) scores between each feature and the target variable

   ii) Rank features based on MI scores
7. Select Features Using MI

   i) Retain features with MI scores above a predefined threshold
8. **Embedded Method:** Tree-Based Feature Selection

   i) Train a tree-based classifier using the filtered feature set

   ii) Compute feature contribution scores using the trained tree-based model

   iii) Select features with importance values above a defined threshold
9. **Evaluate the model performance:**

   i) Calculated Value Accuracy, Precision, Recall, F1-Score

   ii) Generate Confusion Matrix

Output: Selected features evaluation metrics

Figure 1: Hybrid Feature Selection -Filter (Mutual Information) + Embedded (Tree-Based) Method

The findings given in Table 2 and Figure 2 show that, even though all three classification methods have a positive influence on the general performance, the embedded tree-based approach offers a better balance of accuracy, sensitivity, and specificity. This balanced performance contributes to its application in medical diagnostic tasks that most certainly require the detection of the positive cases (sensitivity) and the negative cases (specificity) in an accurate manner.

The advantage of filter based Mutual Information technique is that it is both computationally efficient and fast hence highly applicable in the initial analysis. Recursive Feature Elimination (RFE) is a wrapper algorithm which is more computationally expensive and prone to scaling and generalization since it depends on the underlying classifier. The findings of these studies substantiate the importance of selecting appropriate classification methods to increase the predictive validity in the process of diagnosing gestational hypertension and preeclampsia.

Table 2: Best Performance of Classification Algorithms Based on Systolic and Diastolic BP

Techniques	*Accuracy*	*Precision*	*Recall*	*F1- Score*	*Sensitivity*	*Specificity*
MI Filter Method Classification	0.92	0.93	0.92	0.93	0.92	0.93
RFE-Based Wrapper Classification	0.78	0.81	0.79	0.78	0.88	0.89
Tree-Based Embedded Classification	0.92	0.93	0.94	0.92	0.94	0.94

Table 3 and Fig. 3 are indicative of the performance of the classification in one way or the other. The overall success of the filter based Mutual Information (MI) method is very promising which has a score of 0.92 with equal accuracy, recall, sensitivity and specificity. This means that it is able to establish the influential features at low computation cost, but since it is a filter-based algorithm, it does not consider the interaction between the chosen classifier. RFE algorithm utilizes wrappers on relatively low results on all the evaluation measures with an accuracy of 0.78 and lower precision of 0.81. This decrease in performance can be blamed on the fact that this is a more complex and more prone to over fitting algorithm that limits it in regard to scalability and even to generalization especially when dealing with medical data.

The integrated tree-based algorithm is more effective in recall with a sensitivity and specificity of 0.94 each indicating that it is able to learn nonlinear dependencies among classifier in model training. This increased the diagnosis of positive cases particularly in medical diagnosis. The hybrid system is grounded on the integration between the Mutual Information and tree-based embedded algorithms, which proves to be the most suitable in terms of all the assessment measures including accuracy of 0.98, F1-score of 0.98, sensitivity of 0.97, and specificity of 0.98. All these findings show that hybrid approach is effective in the removal of unnecessary and redundant features and also maintaining all the information that are clinically relevant that makes the classification more efficient and robust.

The comparison and analysis contrasts three traditional ML classification methods - decision tree, random forest and XGBoost against the proposed hybrid model, the combination of Mutual Information (filter method) with a tree-based embedded method. The results of the experiment show that the baseline models performed at an acceptable range. Decision tree and Random Forest both gave an accuracy of 0.90 which indicates similar and valid predictive power. Random Forest among them demonstrated a slightly better overall performance with the best F1score of 0.93, which was associated with a strong precision–recall trade-off. Comparatively, XGBoost had a lower performance with an Accuracy of 0.86 and an F1 score of 0.68. It implies a relatively low capability to balance the false-positives and false-negatives. This sensitivity analysis can further be done to show that Decision Tree had the best recall at 0.93 and this indicates that it has a high ability to recognize the positive cases correctly. At the same time, the recall of Random Forest and XGBoost amounted to 0.85 and 0.84, respectively, which means that they are relatively less effective in identifying positive cases. These results are consistent with earlier studies reported by Nuryunarsih et al.,2025, Amelia et al.,2025, and Amin et al.,2025.

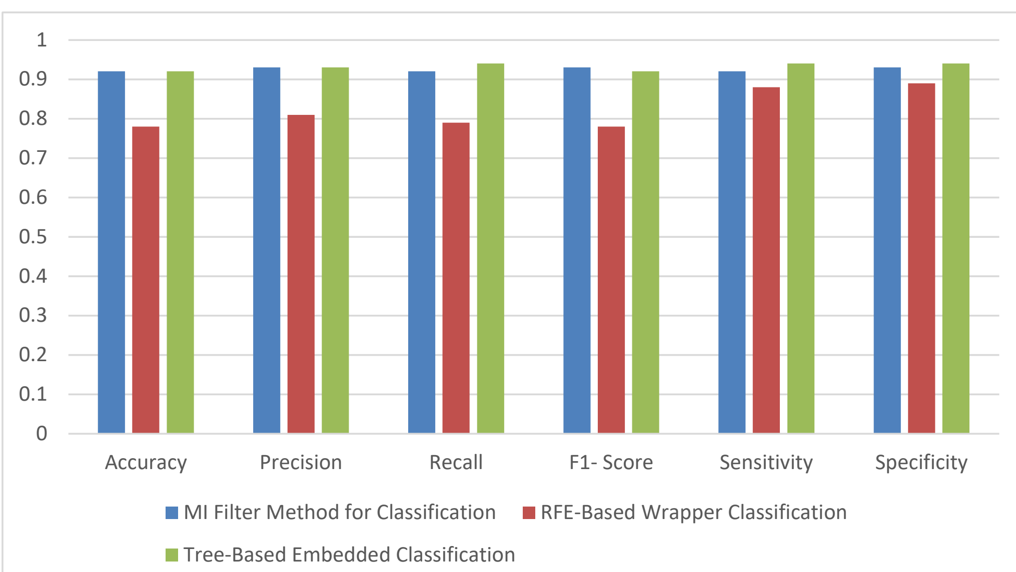

Fig. 2: Best Performance of Classification Algorithms Using SBP and DBP

Table 3: Comparative Analysis of Existing and Proposed (MI + Tree based) Methods

Techniques	*Accuracy*	*Precision*	*Recall*	*F1-Score*	*Sensitivity*	*Specificity*
MI Filter Method for Classification	0.92	0.93	0.92	0.93	0.92	0.93
RFE-Based Wrapper Classification	0.78	0.81	0.79	0.78	0.88	0.89
Tree-Based Embedded Classification	0.92	0.93	0.94	0.92	0.94	0.94
Hybrid Feature Selection - Filter (Mutual Info) + Embedded (Tree-Based) Method	0.98	0.98	0.97	0.98	0.97	0.98

As compared to all the models employed as the baseline, the proposed Hybrid method was much more successful in all the measures of evaluation. It was the most accurate (0.98) and precise (0.98) and f1-score (0.97) and sensitivity (0.98). The good performance shows that, Mutual Information together with a tree based embedded model is capable of identifying the best and informative feature and eliminating redundant ones. This optimized has the advantage of maximizing predictive capability of the classifier, reducing the imbalance of the different types of errors, and generally improving the general model resilience. Therefore, the hybrid structure is more practical and robust than the conventional machine learning approaches, which implies that it is more applicable in predictive modeling with accuracy and efficiency.

Confusion Matrix

It is a structured representation used to evaluate how well a classification model performs by contrasting its predicted categories with the actual outcomes. It also gives a breakdown of the accurate predictions and certain forms of classification error e.g. false positives and false negatives.

It is built with the following values:

True Positives (TP)

Cases in which the occurrence of GH/PE is correctly identified by the model.

True Negatives (TN)

Cases in which patients who do not have GH with PE are properly projected as not.

False Positives (FP)

Cases in which the model incorrectly predicts the presence of GH with preeclampsia in patients who do not have the condition.

False Negatives (FN)

Cases in which the model wrongly forecasts the absence of GH with preeclampsia in patients who actually have the condition.

Accuracy

Accuracy represents the ratio of correctly classified instances amongst all forecasts generated by the model.

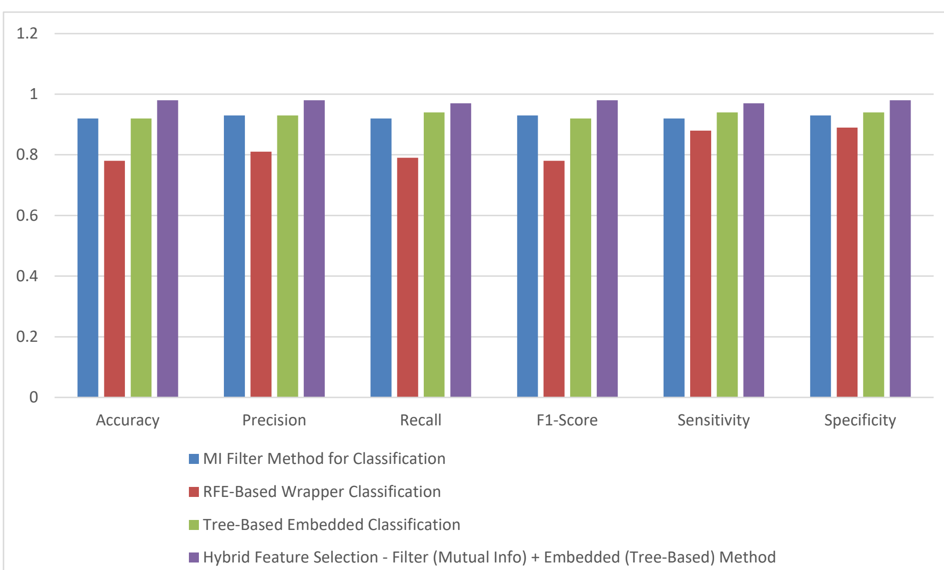

Figure 3: Performance Comparison of Existing and Proposed Hybrid (MI + Tree-Based)

Table 4: Performance Comparison of Existing and Proposed Hybrid (MI+Tree) Methods

Methods	*Decision Tree*	*Random Forest*	*Xgboost*	*Hybrid Feature Selection - Filter (MI) + Embedded (Tree) Method*
Accuracy	0.90	0.90	0.86	0.98
Precision	0.90	0.88	0.87	0.98
F1-score	0.88	0.93	0.68	0.97
Sensitivity	0.93	0.85	0.84	0.98

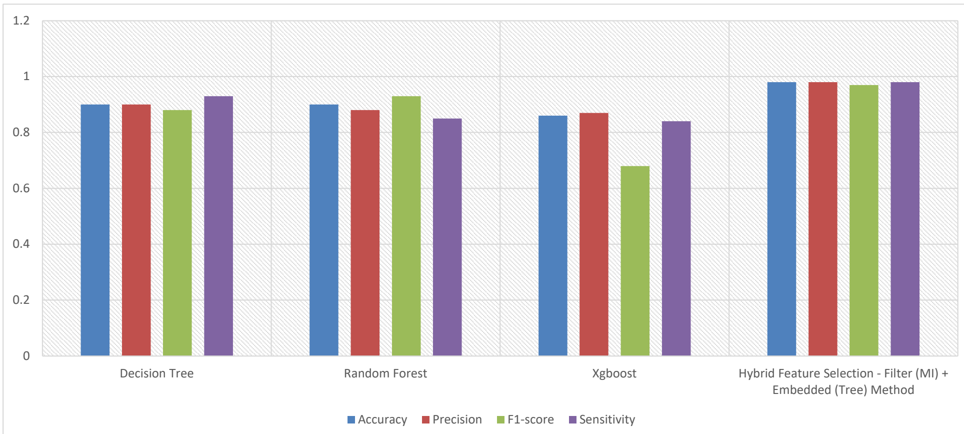

Fig.4: Performance Comparison of Existing vs. Proposed (MI + Tree-Based) Approach

$$\text{Accuracy} = \frac{TN + TN}{TP + FP + FN + TN}$$

Precision

It measures the proportion of true proportion of correctly identified positives among predicted positives.

$$\text{Precision} = \frac{TP}{TP + FP}$$

Recall (Sensitivity)

This metric assesses the model's sensitivity in recognizing positive instances by comparing correctly predicted positives with all actual positive observations.

$$\text{Recall} = \frac{TP}{TP + FN}$$

Specificity

It reflects the model's effectiveness in identifying non-positive cases, expressed as the ratio of correctly predicted negatives to the total number of real negatives.

$$\text{Specificity} = \frac{TN}{TN + FP}$$

F1-score

The F1-score is the harmonic mean of precision and recall. It provides a balanced evaluation of the model's performance, particularly when class distribution is uneven.

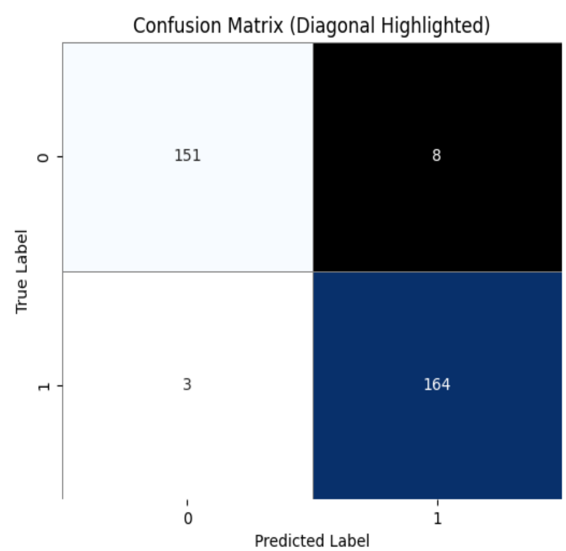

Fig. 5: Confusion Matrix of Existing MI-Based Filter Classification Model

$$F1 - Score = \frac{(2{*}Precision{*}Sensitivity)}{Precision + Sensitivity}$$

As depicted in Figure 5, the confusion matrix of the Filter Method using Mutual Information (MI) classification demonstrates a high level of predictive performance. It was able to identify 151 negative cases as it was expected to, which is an indication that this model can also be used

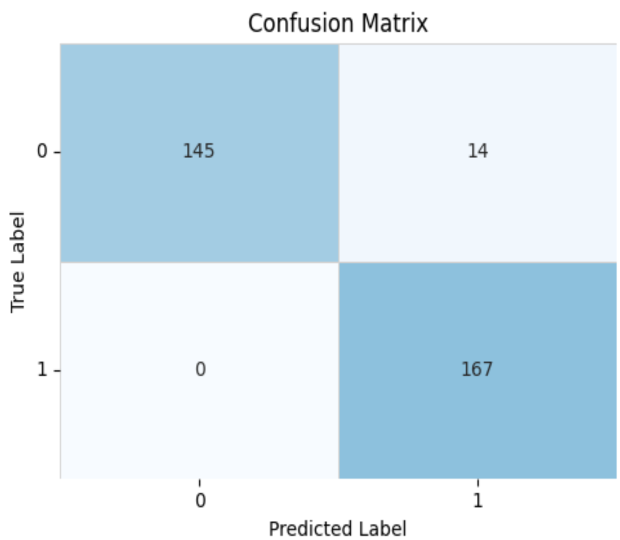

Figure 6: Confusion Matrix of Existing RFE-Based Wrapper Classification Model

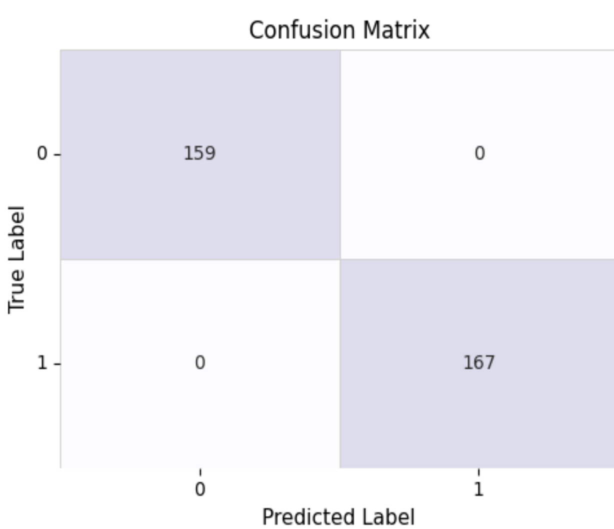

Figure 7: Confusion Matrix Comparison of Existing and Proposed Models

to identify the low-risk cases. It also had 8 false positive prediction. there were few negative cases which were identified as positive and did not impact much on precision. The model also provides 3 false negatives and this implies that the number of positive cases failing was very small hence indicates high sensitivity which is a very important criterion of medical diagnosis. The number of positive cases that were identified was 164 which shows the model's high potential in the detection of the high-risk cases. the results indicate that, the MI-based filter method possesses a high classification accuracy and a low misclassification, therefore, it can be used in medical prediction.

The confusion Table below in Figure 6 measures the effectiveness of the model in classification through four paramount elements, which are TP, TN, FP, FN as explained the model properly recognized 145 negative or low-risk cases and this proves that the model can effectively identify non-preeclampsia cases. The false positive prediction of cases amounted to 14 cases, which mean that there are a few instances of misclassification that has a negligible impact on the specificity. At this point of time, it is essential to notice that the deployed model did not give any false negatives hence all cases with a high-risk factor have been identified successfully and this is important in the diagnosis of medical cases. Also, 167 positive cases were properly identified, which demonstrates the high sensitivity and the high ability of this model to determine preeclampsia conditions. The findings reported show that the model attains good and consistent classification.

In the given Fig.7, the confusion matrix indicates that there is a good predictive performance of the classification model. The model correctly recognized 159 cases that can be categorized as the negative cases which demonstrates that the model is useful in recognizing the negative cases or low-risk cases. The model also had no false positives, that is, no negative cases were mistakenly declared as positive, therefore, having a high degree of precision and reliability. the model showed no false negatives, a fact that assures the absence of any false positive cases and that this is an important factor particularly in the diagnosis of medical conditions. In addition, 167 instances of the positive category were appropriately marked in the model (True Positive), and it demonstrates that the model is highly able to identify high risk.

Conclusion

Preeclampsia and gestational hypertension are severe conditions during pregnancy that creates significant risks to the health of both the mother and the baby when left undetected at an early stage. In this research, the comparative statistical assessment of machine learning classification models established on clinical and physiological parameters was provided with systolic blood pressure (SBP) and diastolic blood pressure (DBP) as the most significant predictors. The Mutual Information (MI) filter approach demonstrated a reasonable level of performance. The accuracy of the method was 0.92, and both the recall, sensitivity, and specificity were balanced. It is a filter-based method, however, so it does not capture the more complex inter-feature interactions in the classifier. RFE wrapper approach was appeared to be relatively less effective (accuracy value of 0.88 and precision of 0.85) possibly because it was more complex to compute and it is prone to overfitting. Comparatively, the tree-based method had a greater level of diagnostic accuracy, with a sensitivity and specificity rate standing at 0.94 and was able to effectively model complicated non-linear dependencies between features.

The hybrid is the combination of Mutual Information and with a tree based embedded classifier and an accuracy of 0.98, an F1-score of 0.98, the sensitivity of 0.97 and specificity of 0.98 were found to be the most effective in the overall performance. These results emphasize the applicability of suggested hybrid classification regarding medical diagnosis The provided system presents a powerful and efficient way of early diagnosing gestational hypertension and preeclampsia and a time-sensitive clinical intervention and improved maternal treatment outcomes under the condition of regular SBP and DBP measurements.

Acknowledgment

The research scholar, Ms. R. Gomathi, shows her heartfelt thanks to her research supervisor, Dr. K. Menaka and the management of Urumu Dhanalakshmi College in support as well as in provision of the resources needed to conduct this study.

References

Ali Raza, & Siddiqui, H. U. R. (2022). Ensemble learning-based feature engineering to analyze maternal health during pregnancy and health risk prediction. PLoS ONE, 17(11), e0276525. https://doi.org/10.1371/journal.pone.0276525

Amelia, D., Adisasmita, A., & Siregar, K. N. (2025). Machine learning for preeclampsia prediction: Enhancing screening in primary health care. Kesmas: Jurnal Kesehatan Masyarakat Nasional, 20(2), 147–156. https://doi.org/10.7454/kesmas.v20i2.2243

Amin, P., & Ahmad, S. G. (2025). Predictive analytics in maternal health: A machine learning approach for classification of preeclampsia. Journal of Maternal-Fetal Medicine. https://doi.org/10.1007/s44196-025-00825-4

Antony Sekar, Taniya, K., & Balamurugan, R. (2025). Enhanced preeclampsia risk prediction with a hybrid machine learning approach. International Journal for Research in Applied Science & Engineering Technology, 13(1).

Bennettand, R., & Mulla, Z. D. (2022). An imbalance-aware deep neural network for early prediction of preeclampsia. PLoS ONE, 17(4), e0266042. https://doi.org/10.1371/journal.pone.0266042

Bora, K., & Kalimuthusamy, N. (2025). A machine learning approach to predict hypertension using cross-sectional and two-year follow-up data from a health and demographic cohort of Assam, North East India. Indian Journal of Medical Research, 161, 394–405. https://doi.org/10.25259/IJMR_881_2024

Devi, S., & Bhagat, P. P. (2024). Novel maternal risk factors for preeclampsia prediction using machine learning algorithms. IAES International Journal of Artificial Intelligence, 13(4), 4544–4556. https://doi.org/10.11591/ijai.v13.i4.pp4544-4556

Dewangan, S., & Dewangan, M. (2025). Retrospective analysis to predict preeclampsia by machine learning algorithms and its relation to neonatal outcome. Indian Journal of Obstetrics and Gynecology Research, 12(3), 540–546. https://doi.org/10.18231/j.ijogr.v.12.i.3.29

Hasan, A. S., & Debu, S. S. S. D. (2024). Machine learning models for predicting risky pregnancies in early clinical interventions. Angiotherapy, 1–11. https://doi.org/10.25163/angiotherapy.889905

Harizahayu, Friendly, & Seputro, B. P. (2024). Predictive modeling of preeclampsia risk using random forest algorithm within a machine learning framework. Journal of Computer Networks, Architecture and High-Performance Computing, 6(4). https://doi.org/10.47709/cnahpc.v6i4.4779

Hedley, P. L., & Hagen, C. M. (2023). The use of artificial intelligence and machine learning methods in early pregnancy preeclampsia screening: A systematic review protocol. PLoS ONE, 18(4), e0272465. https://doi.org/10.1371/journal.pone.0272465

Jonathan, S., & Kadambi, A. (2024). Using machine learning to predict the risk of developing hypertensive disorders of pregnancy using a contemporary nulliparous cohort. Elsevier, Article 100386. https://doi.org/10.1016/j.xagr.2024.100386

Khadidos, A. O., & Saleem, F. (2024). Ensemble machine learning framework for predicting maternal health risk during pregnancy. Scientific Reports, 14, 21483. https://doi.org/10.1038/s41598-024-71934-x

Martínez-Velasco, A., & Martínez-Villaseñor, L. (2018). Machine learning approach for pre-eclampsia risk factors association. Proceedings of Goodtechs ’18, ACM. https://doi.org/10.1145/3284869.3284912

Meshram, J., & Devi, S. (2023). Machine learning algorithm and GESTOSIS score–assisted high-risk pregnancy-induced hypertension prediction. Revue d’Intelligence Artificielle, 37(1), 117–128. https://doi.org/10.18280/ria.370115

Mousavi, S. S. S., & Tierney, K. (2024). Early prediction of hypertensive disorders of pregnancy using machine learning and medical records from the first and second trimesters. IEEE Journal of Biomedical and Health Informatics. https://doi.org/10.1101/2024.11.21.24317720

Nuryunarsih, D., & Rauf, S. (2025). Predicting systolic and diastolic blood pressure response using machine learning: A 96-feature analysis in hypertensive patients with comorbidities. Journal of Current Science and Technology, 15(4), 140. https://doi.org/10.59796/jcst.V15N4.2025.140

Pi, X., & Wang, J. (2025). Prediction of high-risk pregnancy based on machine learning algorithms. Scientific Reports, 15, 15561. https://doi.org/10.1038/s41598-025-00450-3

Shahadat Uddin, Khan, A., Hossain, M. E., & Moni, M. A. (2019). Comparing different supervised machine learning algorithms for disease prediction. BMC Medical Informatics and Decision Making, 19, 1–16. https://doi.org/10.1186/s12911-019-1004-8

Shyu, I., Liu, C.-F., & Tsai, Y.-C. (2025). Machine learning predictive system to predict the risk of developing pre-eclampsia. BMJ Health & Care Informatics, 32, e101151. https://doi.org/10.1136/bmjhci-2024-101151

Tahir, M., Badriyah, T., & Syarif, I. (2018). Classification algorithms of maternal risk detection for preeclampsia with hypertension during pregnancy using particle swarm optimization. EMITTER International Journal of Engineering Technology, 6(2).

Thakur, R., Mansha, & Sharma, P. (2023). Diseases prediction using classification algorithm. International Journal for Research in Applied Science and Engineering Technology, 11(8).

Vinnarasi, P., & Menaka, K. (2025). Predicting the cause of vitamin D deficiency on thyroid and autoimmune disorders using machine learning classifiers. Indian Journal of Natural

Sciences, 16(90).

Wang, C., Zeng, T., & Zhao, X. (2025). Development and validation of a predictive model for preeclampsia: A retrospective cohort study. Archives of Gynecology and Obstetrics, 312, 823–832. https://doi.org/10.1007/s00404-025-08076-6

Yanto, Gusrino, & Puspita, S. (2025). A random forest algorithm for high-risk pregnancies prediction based on explainable artificial intelligence. Communications in Mathematical Biology and Neuroscience. https://doi.org/10.28919/cmbn/9536

Zhang, X., & Chen, Y. (2022). Prediction of severe preeclampsia in machine learning. Elsevier. https://doi.org/10.1016/j.medntd.2022.100158

Zhang, Y., & Gu, X. (2025). Prediction models for late-onset preeclampsia: A study based on logistic regression, support vector machine, and extreme gradient boosting models. Biomedicines, 13, 347. https://doi.org/10.3390/biomedicines1302034.