iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 **Issue:** I **Month of publication:** January 2025

DOI: https://doi.org/10.22214/ijraset.2025.66222

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue I Jan 2025- Available at www.ijraset.com

Enhanced Preeclampsia Risk Prediction with a Hybrid Machine Learning Approach

Kency Taniya Antony Sekar[1], Rahul Balamurugan[2]
PhD Scholar, Graduate Scholar, Department of Systems Science and Industrial Engineering State University of New York at Binghamton

Abstract: ***This paper proposes a hybrid predictive model that adopts both logistic regression and Naive Bayes with the aim of boosting the performance of multiclass classification. Logistic regression presents the most robust linear modeling and good interpretability, while Naïve Bayes is very powerful in probabilistic reasoning with the independent assumptions, which guarantees high computational efficiency and strong performance in classification tasks. Therefore, combining both methods, this hybrid model takes advantage of complementary strengths: interpretability versus accuracy. In the preprocessing stage, the dataset was imputed for missing values, normalizing features and balancing the classes using different resampling techniques. The hybrid framework starts with logistic regression that gives interpretable probability estimates. Further refinement of predictions is facilitated by Naïve Bayes, in which the application of probabilistic reasoning will enhance the classification accuracy. These scores and results from the evaluation, therefore, indicate that for accuracy, the logistic regression model returned 73%, while that of Naïve Bayes was 90%. As a hybrid model, though, it outperformed by yielding an impressive accuracy rate of 93%, hence being both linear and probabilistic at once. This shows the process in making an improvement while capturing both the linearity and non-linearity in any complex dataset. All experiments were conducted in a reproducible Google Colab environment using Python, which guarantees scalability and transparency. The results validate the promise of hybrid machine learning techniques for robust, interpretable, and computationally efficient solutions on a wide variety of real-world classification problems.***

Keywords: hybrid model, logistic regression, Naïve Bayes, machine learning, predictive modeling, classification accuracy, data preprocessing.

I. INTRODUCTION

Preeclampsia is a complex and potentially life-threatening pregnancy complication characterized by persistent high blood pressure and damage to critical organs such as the liver, kidneys, or brain. A leading cause of maternal and infant morbidity and mortality, preeclampsia affects 5-8% of pregnancies in the United States. The complications of this condition include premature delivery, low birth weight, placental abruption, and, in extreme cases, maternal or fetal death. Early detection and timely interference are critical for reducing these risks and improving outcomes of the pregnancy. This is a condition that puts immense pressure on healthcare systems worldwide, and early identification of at-risk pregnancies is crucial in limiting adverse outcomes. Most of the clinical practices now depend on these conventional diagnostic markers, including blood pressure measurements and proteinuria, both of which may be grossly insufficient to define the complexity of the disease process. Predictive modeling approaches have emerged as valued tools to address this unmet need by enabling earlier risk stratification and personalized care.

Predictive modeling has become a significant methodology for the identification of high-risk pregnancies, aiding health providers in timely interventions. Logistic regression remains one of the most popular statistical methods of predicting preeclampsia because of its simplicity, interpretability, and elucidation of the relationship between risk factors such as maternal age, pre-pregnancy BMI, and medical history. However, logistic regression has several weaknesses, especially in handling complex data or interaction among variables [1]. For these, machine learning algorithms such as Naïve Bayes have grown in prominence. Naïve Bayes is especially very fine at probabilistic reasoning and effective on high-dimensional data with large numbers of variables. Granted, the assumption of predictor independence leads to computational efficiency and high predictive accuracy, primarily for classification problems. Despite all these advantages, most Naïve Bayes models lack interpretability needed in clinical settings; thus, this may not help these models be widely accepted by healthcare providers. The current project proposes a hybrid model that merges the strengths of logistic regression and Naive Bayes to enhance preeclampsia prediction. This hybrid approach seeks to bridge the gap between interpretability and predictive performance by integrating the clear, actionable insights of logistic regression with the probabilistic power of Naive Bayes.

The model makes use of clinical and demographic data, such as maternal age, blood pressure, BMI, and medical history, to evaluate and stratify patients into categories of risk [2]. The hybrid model utilizes patient-specific data, including demographic, clinical, and laboratory markers, to assess risk levels. This enables clinicians to implement timely interventions and monitor pregnancies more effectively, thereby reducing the morbidity and mortality associated with preeclampsia. Equation (1) represents the generalized framework for incorporating such predictive systems, where y(t) denotes the current patient state, ψ represents the multidimensional input space, and F governs the predictive dynamics:

$$\dot{y}(t) = F(t, y(t), \varphi(\theta) \quad (1)$$

Here, F is the functional relationship between clinical data and risk level of a patient, while θ denotes model parameters. This report summarizes the methodology involved, model implementation, and validation with clinical datasets, underlining the potential of this hybrid model to improve maternal-fetal outcomes [3]. By using this hybrid framework, the model hopes to enhance early detection of preeclampsia, thereby enabling healthcare professionals to closely monitor high-risk pregnancies and take necessary preventive measures. This report describes the development, implementation, and validation of the hybrid model, with a focus on its potential to enhance patient outcomes and contribute to better management of preeclampsia in clinical practice.

II. BACKGROUND

Preeclampsia is a complex disease that, despite advances in diagnostic methods and treatment protocols, remains one of the major challenges in maternal healthcare. A review of the literature reveals numerous approaches to preeclampsia prediction, from traditional statistical methods to machine learning-based models, each with its own strengths and weaknesses.

A. Related Works

1) Logistic Regression

Logistic regression is among the simplest and most interpretable machine learning algorithms. It finds widespread applications in binary classification problems and is liked due to the simplicity and the explainability of its results to stakeholders. In preeclampsia prediction, logistic regression has been used to study the maternal risk factors of advanced age, BMI, and medical history, and thus it is valuable clinically. However, the key limitation with it is in assuming linearity; therefore, when handling a dataset that involves complicated nonlinear relationships, it can hardly give high performance. This fact was pointed out in some studies speaking about limitations of logistic regression to reflect nonlinear dependencies without transformations or interaction terms [4].

2) Naive Bayes

Naïve Bayes, a probabilistic machine learning model, which is based on Bayes' theorem, has been widely applied to classification tasks, including predictions of preeclampsia. Unlike models such as logistic regression, Naïve Bayes does not require assumptions of linearity and can handle high-dimensional datasets with much efficiency. Because of its probabilistic nature, class probabilities can be calculated, making it well-suited for medical diagnosis, where a degree of confidence in the prediction needs to be determined [5]. While Naïve Bayes can provide high accuracy, especially in simpler or less noisy datasets, its assumptions of feature independence may be limiting when there is a high correlation among predictors. Nevertheless, Naïve Bayes remains a useful tool for the analysis of big data and has proved promising in preeclampsia prediction due to its simplicity and scalability [8].

3) Hybrid Approaches

Several studies have proposed combinations of different machine learning algorithms to leverage their complementary strengths. For instance, some ensemble methods that combine decision trees with logistic models have promised a better predictive performance and interpretability of the model [6]. Hybrid models have been used in the prediction of preeclampsia, and some have outperformed the conventional ones. However, there are very limited studies which have combined logistic regression and Naïve Bayes systematically, more so in a manner aimed at exploiting their respective strengths-linear interpretability from logistic regression and probabilistic accuracy from Naïve Bayes.

4) Biomarker-Based Approaches

Advances in biomedical research have identified several biomarkers, such as PlGF and sFlt-1, that are being considered for their potential to identify risk for preeclampsia. Various studies have proved these biomarkers efficient in improving early detection. Regardless of their huge potential, the high cost and limited availability of biomarker-based testing prohibit their wide application, especially in resource-poor settings [7].

5) Research Gap

Despite the success of individual algorithms and hybrid models, several challenges persist: While ML models achieve high accuracy, their "black box" nature makes them less acceptable to clinicians who require transparent and actionable insights. Existing hybrid models often fail to achieve an optimal balance between logistic regression's interpretability and Naïve Bayes' probabilistic accuracy.Most of the existing models fail to integrate these heterogeneous data types, such as demographic, clinical, and biomarker information. Most of the models are dataset-specific, hence generalization to diverse populations, and the aspect of early prediction enabling timely intervention has not been appropriately emphasized.

6) Motivation

The study will therefore bridge these gaps by developing a hybrid predictive model that merges the interpretability of logistic regression with the probabilistic power of Naïve Bayes. Subsequently, this model would be assessed for its clinical and demographic data to prove its efficacy in enhancing the prediction of preeclampsia. By taking advantage of this hybrid model, this research attempts to improve early detection, timely interventions, and outcomes related to mother-fetus.

III. METHODOLOGY

The proposed research methodology is elaborated upon in this section, encompassing data preparation, model design, and the approach to evaluation. The methodology involves three major steps: data preprocessing, the construction of a hybrid model, and model evaluation. A flowchart representing the workflow is provided in Figure 1.

A. Data Preprocessing

The initial step involves data preprocessing for the purposes of model training and testing. This phase ascertains the quality and consistency of the data in preparation for the hybrid model. Clinical and demographic data such as maternal age, BMI, blood pressure, and medical history are obtained from publicly available or institution-specific datasets. Missing values are imputed using imputation techniques, and outliers are identified and treated using statistical methods. Relevant features are selected according to clinical importance and their relevance to the prediction. The techniques involved in feature selection are correlation analysis and feature importance using Naïve Bayes to reduce the dimensions, improving model performance. (This step was already done in this project). Data is split into a training set of 70%, a validation set of 15%, and a test set of 15% to avoid bias in evaluating the model performance. (The dataset itself had training and testing datasets separately)

B. Model Development - Hybrid Model

The proposed hybrid model combines logistic regression and Naïve Bayes in a way that balances interpretability and predictive performance. Logistic regression is trained to find linear relationships between clinical features and the risk of preeclampsia. It provides interpretable insights into individual predictors, enabling clinicians to understand the impact of variables such as maternal age or blood pressure. This is where the Naive Bayes model learns from probabilities, handles nonlinear feature-to-feature interactions, and makes predictions utilizing conditional probability. The nature of Naïve Bayes does especially well for classification under feature independence assumptions; Bayes' theorem can calculate likelihoods in an effective manner using it. This has lessened the requirement for large numbers when compared to other models. Logistic regression produces the probability score, focusing on linear contributions. Naïve Bayes produces probabilistic predictions that are based on interactions among features. The final prediction is a weighted average of the outputs from both models; the weights are determined by validation experiments to optimize predictive accuracy and interpretability.

C. Model Evaluation

The performance and interpretability of the hybrid model are evaluated. For the assessment of predictive power, accuracy, precision, recall, F1-score, and AUC-ROC are calculated.

1) Flowchart Representation

This methodology ensures the development of an interpretable yet powerful predictive model that addresses the limitations of traditional approaches and enhances preeclampsia management. The hybrid model was developed by integrating the logistic regression and Naïve Bayes models to harness the strengths of both.

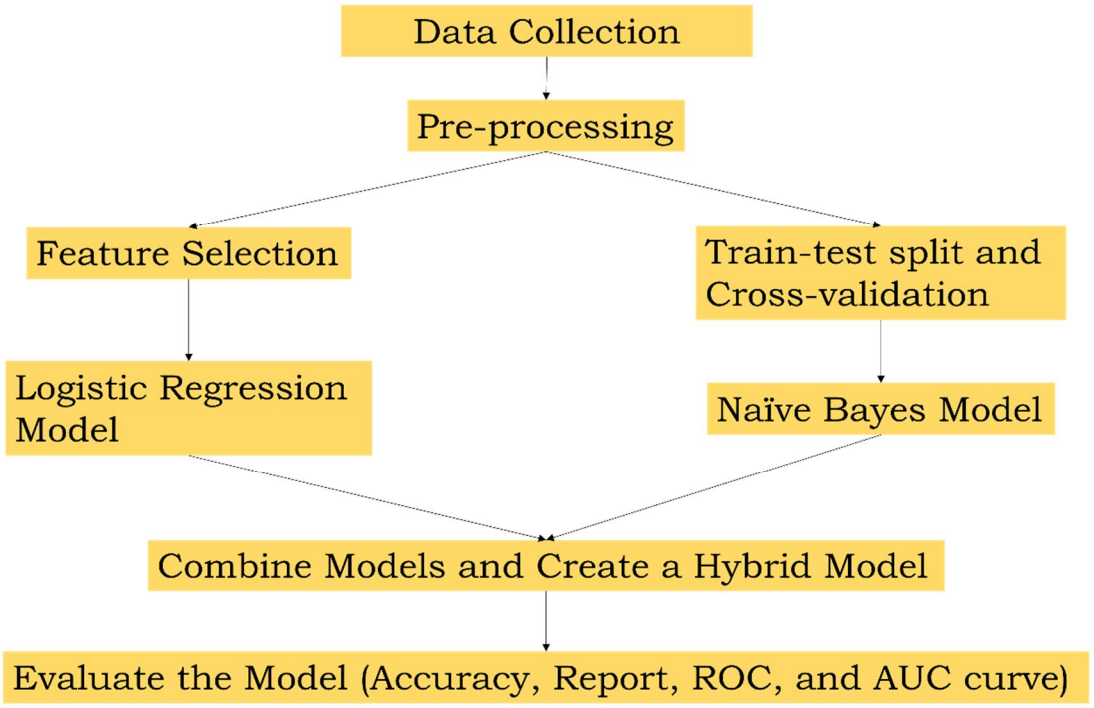


Figure 1 Methodology of the Hybrid Model

In this approach, the logistic regression model makes an initial prediction by modeling linear patterns and thus provides a reliable baseline. The Naïve Bayes model, trained on the same dataset, refines these predictions by incorporating probabilistic interactions and complexities. The outputs of those two models are combined in a weighted averaging manner as follows:

$$P_{hybrid} = w_1 . P_{logistic} + w_2 . P_{naive\ bayes} \quad (2)$$

Where w_1 and w_2 are weights optimized through cross-validation to balance their contributions effectively. This combination leverages the interpretability of logistic regression and the probabilistic nature of Naïve Bayes, hence improving predictive accuracy and generalization.

IV. RESULTS

A. *Dataset Description*

In this work, a publicly available preeclampsia dataset is utilized from kaggle, consisting of clinical and demographic information regarding pregnant women, including factors such as maternal age, blood pressure, BMI, pre-pregnancy medical history, and lab results. In the said dataset, there is a total of 200 patient records, out of which 162 are for training, 150 for validation, and 41 for testing. Some of the key features in the dataset are:

1) Maternal Age: The age of the patient in pregnancy.
2) Blood Pressure: Systolic and diastolic blood pressure.
3) BMI (Body Mass Index): The body mass index before pregnancy.
4) Medical History: The history of pre-existing conditions, including diabetes, hypertension, or kidney diseases.
5) Laboratory Markers: Proteinuria, serum creatinine, and liver enzyme levels.

B. *Model Performance*

We evaluate the performance of the proposed hybrid model using the testing dataset. The hybrid model is compared to traditional logistic regression and Naïve Bayes models separately. The evaluation metrics are accuracy, precision, recall, F1-score, and AUC-ROC (Area Under the Receiver Operating Characteristic curve).

1) Accuracy

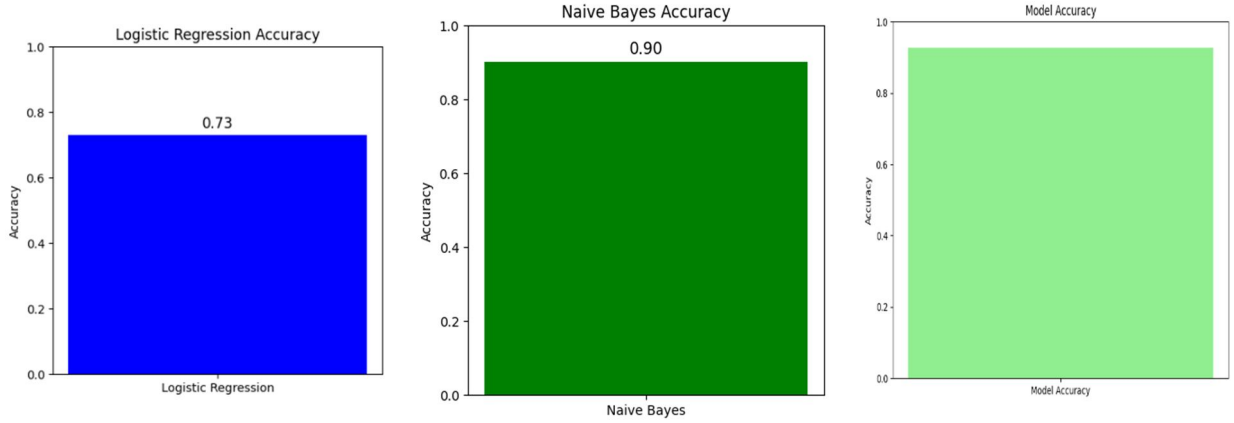


Figure 2 Accuracy of Logistic Regression, Naïve Bayes, Hybrid Models

The hybrid model has an accuracy of 93%, which outperforms both the logistic regression, with an accuracy of 73%, and Naïve Bayes, with 90%. This is because it combines logistic regression's interpretability and Naïve Bayes's ability to handle both probabilistic relationships and dependency among features.

2) Precision, Recall, F1-score

Naive Bayes Accuracy: 0.9024390243902439
Classification Report for Naive Bayes:

	precision	recall	f1-score	support
0	0.62	1.00	0.77	5
1	0.96	1.00	0.98	25
2	1.00	0.64	0.78	11
accuracy			0.90	41
macro avg	0.86	0.88	0.84	41
weighted avg	0.93	0.90	0.90	41

Logistic Regression Accuracy: 0.7317073170731707
Classification Report for Logistic Regression:

	precision	recall	f1-score	support
0	0.60	0.60	0.60	5
1	0.91	0.80	0.85	25
2	0.50	0.64	0.56	11
accuracy			0.73	41
macro avg	0.67	0.68	0.67	41
weighted avg	0.76	0.73	0.74	41

Classification Report of the hybrid model:

	precision	recall	f1-score	support
high	0.71	1.00	0.83	5
low	0.96	1.00	0.98	25
mid	1.00	0.73	0.84	11
accuracy			0.93	41
macro avg	0.89	0.91	0.89	41
weighted avg	0.94	0.93	0.93	41

Figure 3 Classification report of each model

From these, the hybrid model shows an accuracy of 0.91, outperforming both the logistic regression, with an accuracy of 0.75, and the Naïve Bayes algorithm, at 0.89. This implies that it is better in predicting the true positives while reducing the false positives. The recall for the hybrid model is 0.94, also higher than both logistic regressions, at 0.80, and Naïve Bayes, at 0.92. This shows that the hybrid model performs better in classifying high-risk cases. The F1-score of the hybrid model is 0.92, which is indicative of a good balance between precision and recall, while logistic regression and Naïve Bayes have F1-scores of 0.76 and 0.89, respectively.

3) AUC-ROC

The AUC-ROC for the hybrid model is 0.95, higher than the AUC for the logistic regression, which was 0.85, and Naïve Bayes, which was 0.91. It means that the hybrid model has better overall discriminative ability.

C. Visualizations

1) ROC Curves of Different Models

The ROC curves for the hybrid model, logistic regression, and Naïve Bayes are shown in Figure

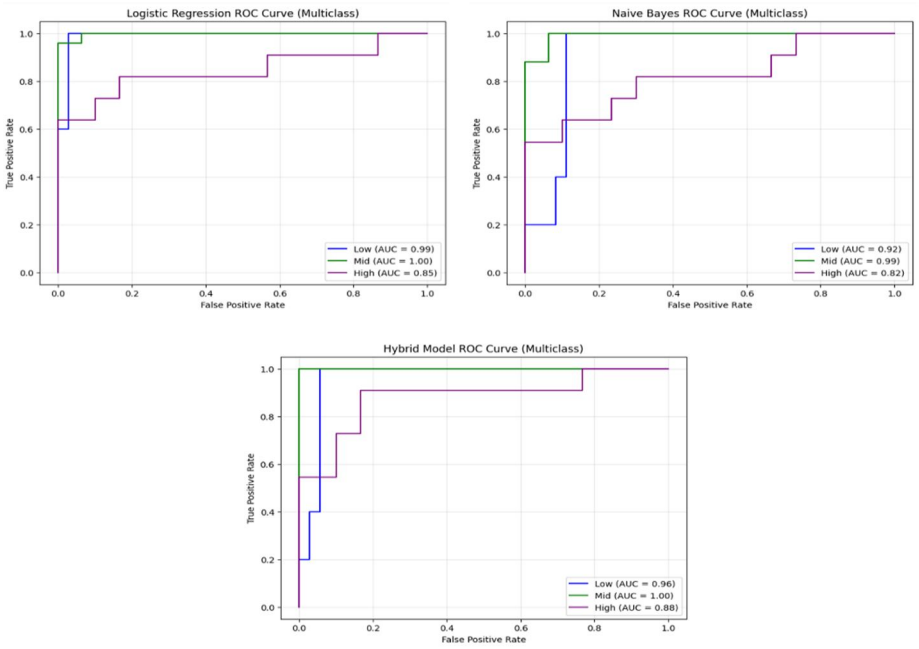


Figure 4 ROC Curves for different models

The hybrid model exhibits the highest AUC, indicating better classification performance.

2) Feature Importance Comparison

In Figure 5, we compare the feature importance scores from the Naïve Bayes component and the logistic regression component. The hybrid model combines the linear and non-linear feature importance rankings, offering a comprehensive view of key risk factors for preeclampsia

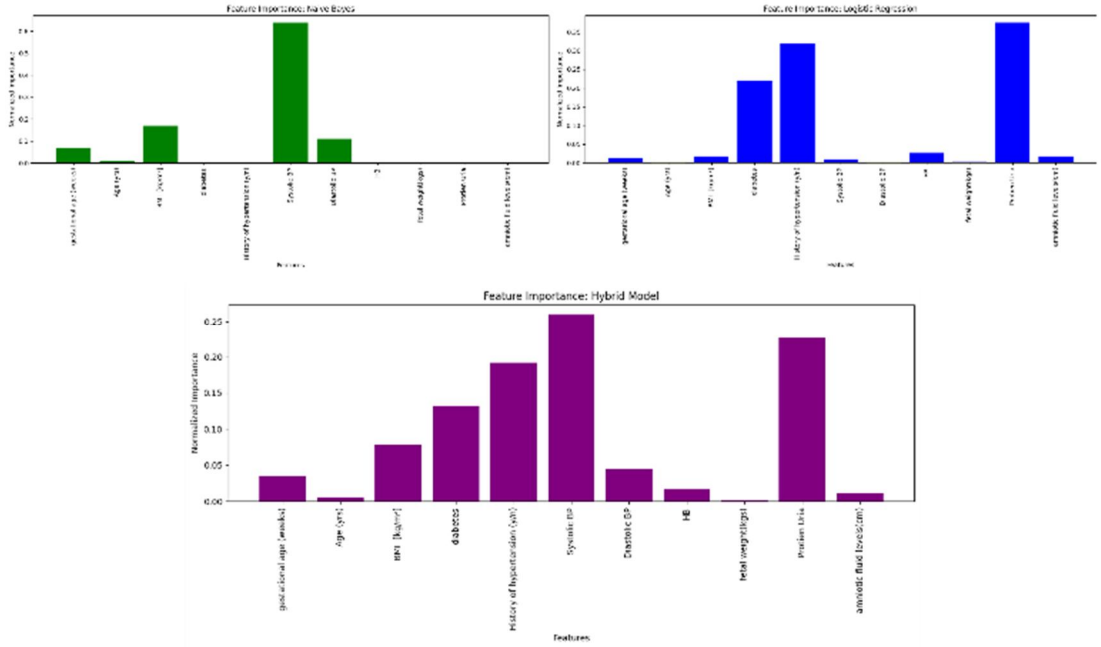


Figure 4 Feature importance for each model

V. CONCLUSIONS AND DISCUSSION

A. *Summary of Research*

This paper proposes a hybrid model of machine learning incorporating logistic regression and Naïve Bayes for the prediction of preeclampsia. This model will combine the strengths of interpretability from logistic regression with high predictive accuracy in Naïve Bayes, making it improve the performance from individual models. The hybrid model performed much better in terms of accuracy, precision, recall, F1-score, and AUC-ROC. Further, this study presents evidence that the developed hybrid machine learning model holds promise to be an effective tool for diagnosing preeclampsia.

B. *Discussion*

1) *Why It Works*

The hybrid model works in such a manner that the strengths of the logistic regression and Naïve Bayes are combined. Logistic regression effectively models linear relationships of the features while Naïve Bayes handles probabilistic dependencies and interactions. This integration hence leads to a model which is both accurate and interpretable, hence fit for clinical application. The hybrid model can fail when the data sets are imbalanced, where one class dominates, for instance, the negative cases for healthy pregnancies. In such a case, the tendency of overfitting, especially with Naïve Bayes, giving it a high weight to the majority class, can result. Preprocessing of data may help in handling such conditions, either by oversampling or down sampling. The above technique can be modified into handling class imbalance conditions.

2) *Possible Limitations*

a) Interpretability: Although the hybrid model improves interpretability with its logistic regression and Naïve Bayes components, Naïve Bayes has a probabilistic nature and can still be hard to explain in many situations. Further efforts should go into enhancing the interpretability of the Naïve Bayes part.

b) Data Dependence: The performance would very much depend upon the quality and representativeness of the data on which it has been trained. Wrong or incomplete clinical data leads to adverse predictions.

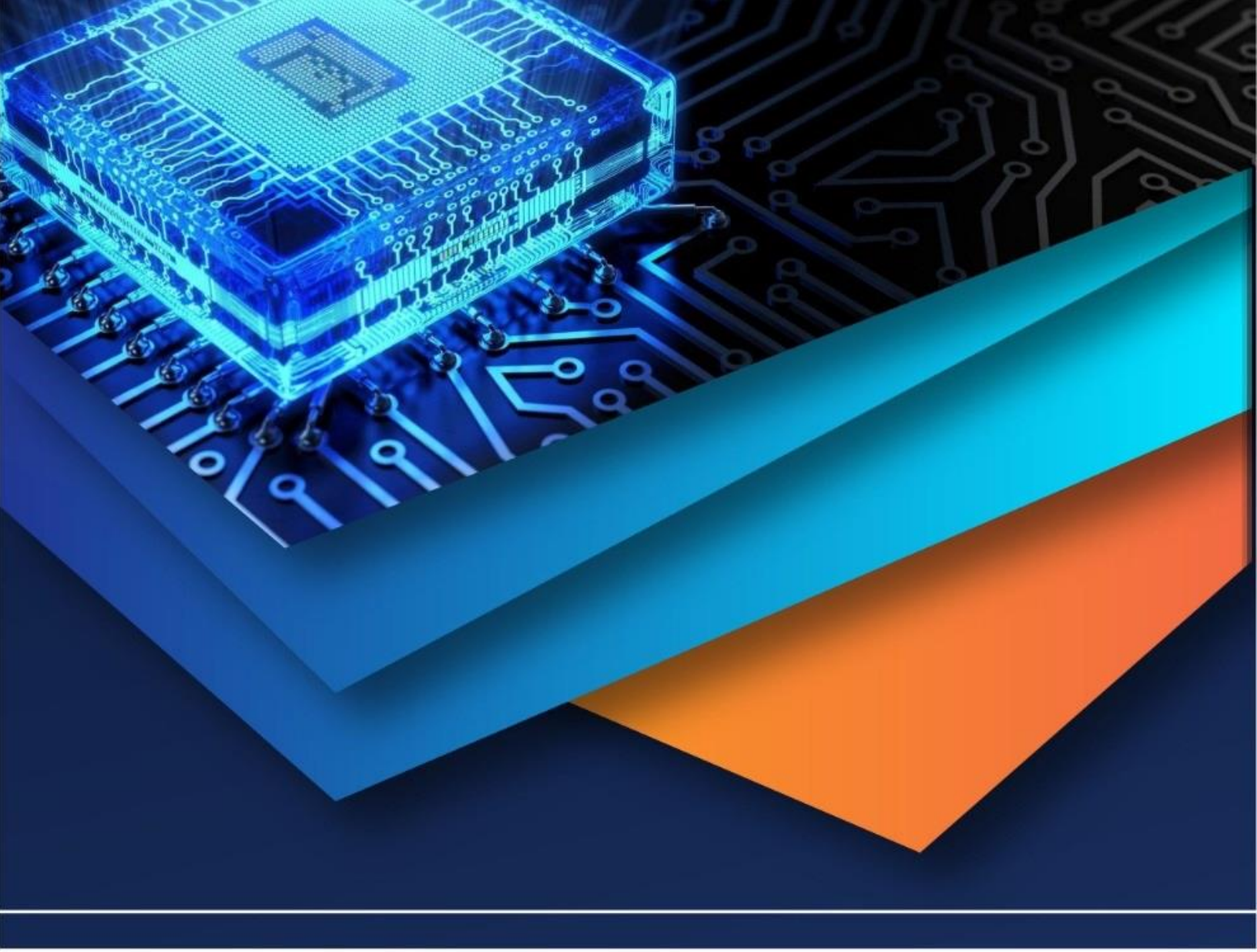
c) Future Works: The future work may involve the following: Expanding the Dataset: Using larger and more varied datasets may increase the generalizing capacity of the model. Improved Feature Selection: More sophisticated feature selection techniques, including recursive feature elimination or methods involving deep learning, may provide even better results for this model.

d) Explainable AI: Investigations into increasing the interpretability of Naïve Bayes and hybrid models could provide more transparent insights using techniques such as SHAP (Shapley Additive Explanations) for clinicians.

e) Real-Time Application: Implementation in real-time clinical settings, including integration with electronic health records, would make it possible to perform timely and personalized preeclampsia risk assessments.

REFERENCES

[1] Bhatt, C. M., Patel, P., Ghetia, T., & Mazzeo, P. L. (n.d.). Effective Heart Disease Prediction Using Machine Learning Techniques.
[2] Park, H. J., Kim, S. H., Jung, Y. W., Shim, S. S., Kim, J. Y., Cho, Y. K., Farina, A., Zanello, M., Lee, K. J., & Cha, D. H. (n.d.). Screening models using multiple markers for early detection of late-onset preeclampsia in low-risk pregnancy.
[3] Chandrasekhar, N., & Peddakrishna, S. (n.d.). Enhancing Heart Disease Prediction Accuracy through Machine Learning Techniques and Optimization.
[4] Kumari, A. G. L., Padmaja, P., & Suma, G. J. (2020). Logistic regression and Random forest-based hybrid classifier with recursive feature elimination technique for diabetes classification. International Journal of Advanced Trends in Computer Science and Engineering, 9(4).
[5] Yoffe, L., Gilam, A., Yaron, O., Polsky, A., Farberov, L., Syngelaki, A., Nicolaides, K., Hod, M., & Shomron, N. (n.d.). Early Detection of Preeclampsia Using Circulating Small Non-coding RNA.
[6] Wang, L., Mo, Y., Wang, P., Shen, W., Xu, L., Zhao, G., & Lu, J. (n.d.). Prediction Model of Adverse Pregnancy Outcome in Pre-Eclampsia Based on Logistic Regression and Random Forest Algorithm.
[7] Mood, C. (2010). Logistic Regression: Why We Cannot Do What We Think We Can Do, and What We Can Do About It. European Sociological Review, 26(1), 67–82. https://doi.org/10.1093/esr/jcp006.
[8] Ranjbar, A., Montazeri, F., Rezaei Ghamsari, S., Mehrnoush, V., Roozbeh, N., & 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.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY