




Optimized Ensemble Learning Based on Genetic Algorithms and R-Shiny App for Early Detection of Preeclampsia



Dina Tri Utari^{*}, Rahmadi Yotenka^{*}, Paringga Fakhri Ashim^{*}

Department of Statistics, Universitas Islam Indonesia, Yogyakarta 55584, Indonesia

Corresponding Author Email: dina.t.utari@uii.ac.id

Copyright: ©2024 The authors. This article is published by IIETA and is licensed under the CC BY 4.0 license (<http://creativecommons.org/licenses/by/4.0/>).

<https://doi.org/10.18280/mmep.111226>

ABSTRACT

Received: 8 September 2024

Revised: 4 November 2024

Accepted: 10 November 2024

Available online: 31 December 2024

Keywords:

preeclampsia, Random Forest, Genetic Algorithm, risk factors, R-Shiny

Early detection of preeclampsia, a potentially life-threatening pregnancy complication, is critical for maternal and fetal well-being. This study offers an innovative approach to early detection using optimized ensemble learning and an interactive R-Shiny application. Genetic Algorithms are employed to enhance the performance of ensemble models—Random Forest. By optimizing model hyperparameters, feature selection, and the ensemble combination, we achieve superior predictive accuracy. The optimal values for the *mtry* parameter of Random Forest are 4 and *ntree* of 9. This outcome was achieved by setting the critical number of individuals to survive at each generation to 2, with a crossover probability of 0.8 and a mutation probability of 0.1. Additionally, an R-Shiny application is developed to provide healthcare professionals with an accessible tool for risk assessment and early intervention. The combination of Genetic Algorithms and ensemble learning, complemented by a user-friendly interface, offers a promising solution for timely preeclampsia diagnosis and proactive healthcare management.

1. INTRODUCTION

Pregnant women diagnosed with preeclampsia exhibit a heightened vulnerability to significant health complications, which encompass elevated blood pressure and the potential for organ impairment, particularly affecting the renal and hepatic systems. A documented history of tobacco usage, alcohol intake, and pre-existing hypertension constitutes contributory risk factors correlated with an augmented probability of developing preeclampsia. The early detection of preeclampsia is paramount for prompt intervention and management, thereby safeguarding the health and well-being of both the mother and the neonate.

A multitude of studies have used machine learning to improve the prediction of preeclampsia. This research underscores the efficacy of advanced algorithms, such as Random Forest, Gradient Boosting, Support Vector Machines, and other machine learning techniques in discerning intricate patterns within patient data that classic statistical methods may overlook.

Jhee et al. [1] investigated the predictive outcomes of late-onset preeclampsia after 34 weeks of gestation. This study sought to create machine learning models for predicting late-onset preeclampsia utilizing hospital electronic medical record data. Logistic Regression, Decision Tree Model, Naive Bayes Classification, Support Vector Machine, Random Forest Algorithm, and Stochastic Gradient Boosting method were employed to develop the predictive models. Integrating maternal variables and prevalent antenatal laboratory data from the early second to the early third trimester can

effectively forecast late-onset preeclampsia utilizing machine learning methods.

The research conducted by Tahir et al. [2] employs Neural Networks to classify preeclampsia data. The outcome indicates a proper classification rate of 96.66% for preeclampsia cases utilizing all factors in the test set. They compare with techniques such as Naive Bayes, K-Nearest Neighbors, Linear Regression, Logistic Regression, and Support Vector Machine. They examined the identification of preeclampsia utilizing a neural network and assessed the significance of the Previous PE Case attribute on the classification outcomes. The experiment demonstrated that the neural network algorithm attained optimal accuracy using three validation methods: 92.46% for split data, 94.23% for 10-fold cross-validation, and 96.66% for leave-one-out (LOO) validation.

Another study suggests employing a machine learning technology called Support Vector Machine for pattern recognition in a pregnancy database. This research presents a thorough inference method for mobile decision support systems that can improve care for women at risk of pregnancy-related complications. This study used the 10-fold cross-validation method to evaluate the suggested model's performance. The advancement of knowledge discovery in health databases has been consistent [3].

The study conducted by C Ramdhani et al. [4] aims to predict pregnancy risk levels through early detection with the Random Forest classification approach, optimized by a Genetic Algorithm. The Maternal Health Risk dataset analysis revealed that Random Forest attained an accuracy of 73.37%,

which significantly improved to 90.20% following Genetic Algorithm optimization. The t-test result confirmed a significant enhancement. These findings indicate that this strategy can efficiently assist medical professionals in evaluating pregnancy health risks and establish a basis for future program development.

Using the Saudi Arabia used cars dataset obtained from the Syarah platform on Kaggle, another research attempt studies the application of machine learning algorithms to anticipate the values of pre-owned autos from the perspective of the Saudi Arabian government. With the help of the model, buyers and sellers will be able to estimate the prices of the automobiles they deal with more correctly. Linear Regression, Random Forest, and XGBoost were the three separate machine-learning techniques investigated. The Mean Squared Error (MSE) scores obtained were 0.15, 0.10, and 0.19 for each of these methodologies. Regarding all evaluative criteria, including RMSE, MSE, and R-squared, the Random Forest Regressor demonstrated exceptional performance [5].

Research findings substantiate the efficacy of machine learning methodologies, notably the ensemble learning framework—Random Forest, in forecasting occurrences of preeclampsia. Utilizing a Genetic Algorithm to optimize Random Forest parameters significantly augments predictive accuracy. This investigation amalgamates these advantageous features, showcasing the optimized model via an intuitive R-Shiny interface, facilitating accessible and practical implementation for healthcare practitioners. Ensemble learning embodies a formidable methodology integrating multiple models to improve predictive reliability and accuracy. Prominent instances of ensemble learning techniques include Random Forest, Gradient Boosting, and Extreme Gradient Boosting. This approach has been thoroughly employed across diverse domains, notably within the healthcare industry.

Random Forest, a machine learning technique that operates by constructing multiple decision trees and aggregating their outputs, is particularly effective in medical diagnostics. Highlights the advantages of Random Forests, including their ability to handle large datasets and their robustness against overfitting, which is especially beneficial in the context of complex medical data [6]. The method's inherent capacity to assess feature importance also aids in identifying critical predictors of preeclampsia, thereby enhancing the interpretability of the model.

Genetic Algorithms are optimization methods derived from the principles of natural selection. They are especially beneficial for feature selection and parameter optimization in machine learning models. Present a detailed examination of Genetic Algorithms, highlighting their utility in optimizing intricate situations where conventional approaches may prove inadequate [7]. Genetic Algorithms emulate the mechanisms of natural selection and evolution to identify the optimal configuration of model parameters [8]. Genetic Algorithms efficiently navigate the solution space and determine the ideal parameters for the Random Forest model by iteratively picking, recombining, and modifying viable solutions. This optimization procedure improves the model's capacity to differentiate between preeclampsia and normal pregnancies, resulting in more precise forecasts.

Recent studies have demonstrated the efficacy of machine learning-based methods, including Random Forest, in predicting preeclampsia and conducted a prospective study

that highlighted the superior predictive performance of machine learning techniques over conventional screening strategies, underscoring the potential of Random Forest in this domain [9]—furthermore, reported high C-statistics for Random Forest models in predicting late-onset preeclampsia, indicating its reliability as a predictive tool [1]. These findings suggest that Random Forests could significantly enhance early detection efforts when optimized through Genetic Algorithms.

The combination of these methodologies improves predictive accuracy and allows for the exploration of complex interactions among various risk factors associated with preeclampsia. It noted that machine learning algorithms, including Random Forest, have shown promise in prognostic prediction studies related to pregnancy care [10]. This highlights the relevance of employing advanced computational techniques to address the multifaceted nature of preeclampsia risk factors.

Moreover, applying Genetic Algorithms to optimize random forest parameters can improve model performance. It emphasized the importance of leveraging biophysical and biochemical markers in predictive modeling for preeclampsia, suggesting that a machine-learning approach could enhance the identification of at-risk individuals [11]. This aligns with the broader trend of utilizing machine-learning to analyze large datasets, as noted by those who highlighted the potential of these algorithms in identifying diagnostic markers for preeclampsia [12].

R-Shiny, an open-source framework for web applications developed in R, facilitates the creation of interactive web applications directly from R scripts, rendering them accessible to individuals with limited programming expertise [13]. Within the public health and epidemiology domain, a Shiny application was developed for small-area disease mapping about cancer incidence, which empowers epidemiologists to assess risks and interactively visualize data [14]. This particular application is a pertinent illustration of how Shiny can augment the analytical capacities of public health practitioners, thereby fostering more informed decision-making grounded in spatial data.

An R-Shiny app has been developed to facilitate the practical application of the optimized Random Forest model. R-Shiny is a web application framework that allows for the creation of interactive and user-friendly interfaces. The development of the R-Shiny app is inspired by the work of Wang and Rhemtulla [15], who utilized R-Shiny for power analysis in structural equation modeling. The app provides healthcare professionals with a user-friendly platform to input patient data and obtain real-time predictions of preeclampsia risk. Integrating the optimized Random Forest model based on the Genetic Algorithm with the R-Shiny app enables efficient and accessible early detection of preeclampsia, potentially improving prenatal care and outcomes.

The novelty of this research lies in the combination of Random Forest and Genetic Algorithms for the early detection of preeclampsia, along with the development of a user-friendly R-Shiny app for real-time predictions and early intervention. While Random Forest has been widely used in various domains, including medical diagnosis, integrating Genetic Algorithms with Random Forest for preeclampsia detection is a novel approach [16-18]. Genetic Algorithms provide an optimization technique that can enhance the performance of the Random Forest model by searching for the optimal set of parameters [16-18].

2. MATERIALS AND METHODS

This study used secondary data sourced from the dataverse.harvard.edu website, which is openly licensed for unrestricted use. The dataset chosen for analysis consisted of replication data pertaining to the determinants of preeclampsia in women who delivered at county hospitals in Nairobi, Kenya [19]. A total of 352 postnatal ward mothers participated in the study, providing relevant information through structured interview questionnaires and the review of their medical records. The dataset was divided into two distinct categories: case of preeclampsia and normal, enabling the creation of sub-datasets for training and testing purposes. The proportions of these sub-datasets were adjusted to optimize the model's performance and yield the most accurate results.

The stages in this research are seen as follows in Figure 1. The initial step is gathering pertinent medical data to provide the basis for further analysis. In the second stage, a Genetic Algorithm is introduced to find the optimized parameter and then apply this best parameter to the Random Forest algorithm. Inspired by biological evolution, Genetic Algorithms adjust the model's parameters to adapt and perform better continuously. The third step involves predicting and achieving the accuracy of classification. The research culminates in developing an interactive R-Shiny application that offers healthcare practitioners an intuitive interface to input patient data and quickly and accurately assess the patient's risk for preeclampsia.

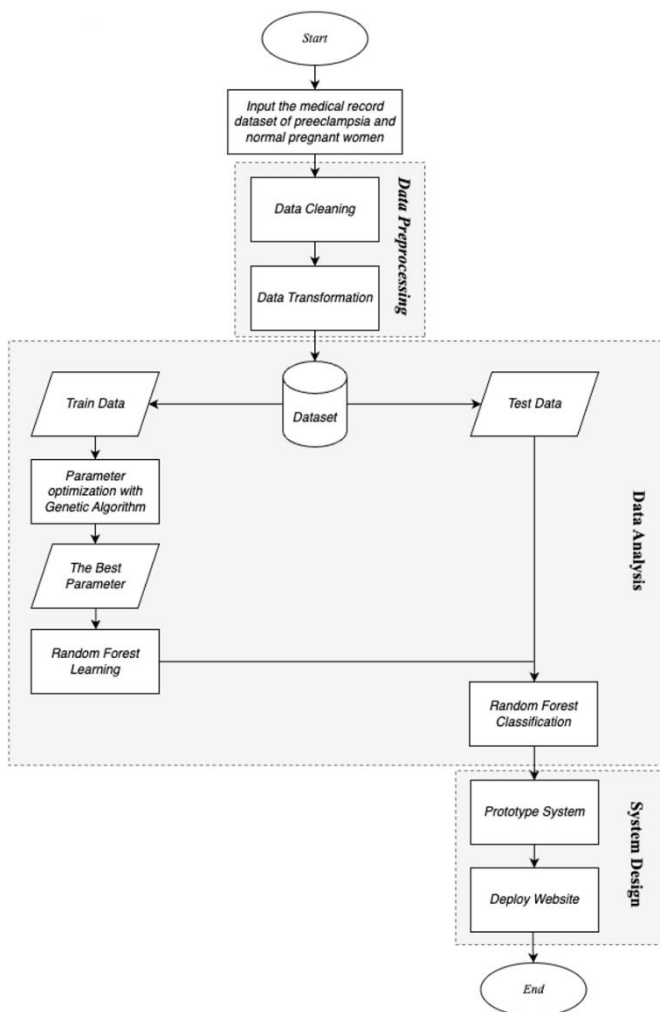


Figure 1. Research stages

2.1 Random Forest

Random Forest represent a significant adaptation of the bagging technique, involving the creation of a diverse ensemble of uncorrelated trees, which are subsequently combined through averaging [20]. A random forest classifier is a meta-estimator that employs decision tree classifiers fitted on various subsets of the dataset. It uses averaging to enhance predictive accuracy and mitigate the risk of overfitting [21]. To mitigate the issue of overfitting, it is imperative to augment the number of decision trees; an elevated number of trees enhances predictive accuracy [22].

Like the name suggests, it is a Random Forest is a tree-based collection of trees, each one dependent on a random collection of variables. More formally, for a p -dimensional random vector $X = (X_1, X_2, \dots, X_p)^T$ representing the real-valued input or predictor variables and a random variable Y representing the real-valued response, we assume unknown joint distribution $P_{XY}(X, Y)$. It is the aim of finding the prediction function $f(X)$ for the prediction of that Y . The prediction function is derived using a loss function $L(Y, f(X))$ and is defined to reduce the cost of the loss,

$$E_{XY}(L(Y, f(X))) \quad (1)$$

in which the subscripts indicate an expectation in relation to the distribution joint of X and Y [23].

The smallest $E_{XY}(L(Y, f(X)))$ to minimize squared error loss provides the expectation of a conditional

$$f(X) = E(Y|X = x) \quad (2)$$

also referred to also regression function.

In the classification situation, if the set of possible values of Y is denoted by \mathcal{Y} , then minimizing $E_{XY}(L(Y, f(X)))$ for zero-one loss gives,

$$f(X) = \operatorname{argmax}_{y \in \mathcal{Y}} P(Y = y|X = x) \quad (3)$$

also referred to by or the Bayes rule.

In which there are K classes that are denoted as $1, 2, \dots, K$ typically, a criteria for splitting is the Gini index,

$$Q = \sum_{k \neq k'}^K \hat{p}_k \hat{p}_{k'} \quad (4)$$

in which \hat{p}_k is the proportion of k class observations in the node:

$$\hat{p}_k = \frac{1}{n} \sum_{i=1}^n I(y_i = k) \quad (5)$$

Random Forest method is an extension of the Classification and Regression Trees (CART) method that incorporates the techniques of bootstrap aggregating (bagging) and random feature selection. In a Random Forest, many trees are grown to create a forest, and then an analysis is conducted on this ensemble of trees [24]. When applied to a dataset with n observations and p explanatory variables, the random forest algorithm operates similarly [25].

1. Bootstrap Sampling: Randomly select n observations from the dataset with replacement, forming a bootstrap sample. This process is known as the bootstrap stage, where each sample is equally likely to be selected.
2. Tree Construction: Using the bootstrap sample, grow a decision tree to its maximum size without pruning. At each tree node, a random subset of m explanatory variables (where m is much smaller than the total number of variables, p) is chosen. From this subset, the best variable for splitting is determined. This stage is referred to as the random feature selection stage.
3. Repeat the Process: Repeat steps 1 and 2 for k iterations, creating a forest of k trees. Each tree is built on a different bootstrap sample, resulting in a diverse ensemble of trees.

2.2 Genetic Algorithm

Genetic Algorithms are optimization algorithms that mimic the process of natural evolution [26]. Inspired by the principles of survival of the fittest, selection, and mutation, Genetic Algorithms simulate the evolution of solutions over multiple generations. The objective is to identify an optimal or close-to-optimal solution for a given optimization problem. The inherent flexibility of genetic algorithms, resembling the process of natural selection, enables these algorithms to modify themselves in response to diverse environments to ensure survival and procreation. A generation within the framework of a genetic algorithm constitutes a collective of entities, and each is engaged in the quest for a solution within the solution space. These entities may be depicted as sequences of binary digits, integers, and similar representations, with each sequence symbolizing a chromosome [27].

A key feature of Genetic Algorithms is their capacity to conduct a worldwide search within the hyperparameter space. In contrast to gradient-based optimization approaches that may settle at local optima, Genetic Algorithms employ principles inspired by natural selection, including selection, crossover, and mutation, to investigate a broad spectrum of solutions [28]. This attribute is especially advantageous for intricate models such as Random Forests, where the hyperparameter landscape may exhibit significant non-linearity and several modes. Although Genetic Algorithms can effectively explore the hyperparameter space, they may also be computationally demanding, particularly with large populations and multiple generations. This computing demand may restrict its usefulness in resource-constrained circumstances [29].

The initial population of potential solutions is determined randomly or through heuristic methods. Each solution is then assessed based on a predetermined fitness function. The individuals with the highest fitness are selected as the best candidates. These top individuals evolve and generate a new population through genetic operators like crossover and mutation. This process continues, evaluating and evolving the population until a termination condition based on the fitness function or the maximum number of generations is reached. Finally, the best individuals from the population are identified and presented as the solution to the optimization problem [30].

In this article the process of selection occurs following mutation and crossover. The fitness function F can be utilized to determine the best individuals. This is defined by the Eq. (6) [31].

$$F_{i,j} = acc_{i,j} - \gamma(L_e - L_s) \quad (6)$$

where, $acc_{i,j}$ is the accuracy for the j -th individual of the generation of i -th obtained through test using the CNN model. The weight is γ , which represents the layers.

The fitness quality is the determining factor in determining whether one who has the highest fitness is chosen as the winner. This study follows the Eq. (7).

$$P_{i,j} = \frac{F_{i,j}}{\sum_{j=1}^M F_{i,j}} \quad (7)$$

where, $P_{i,j}$ is the likelihood that j -th person will survive.

The following outlines the procedure followed by the Genetic Algorithm (Figure 2).

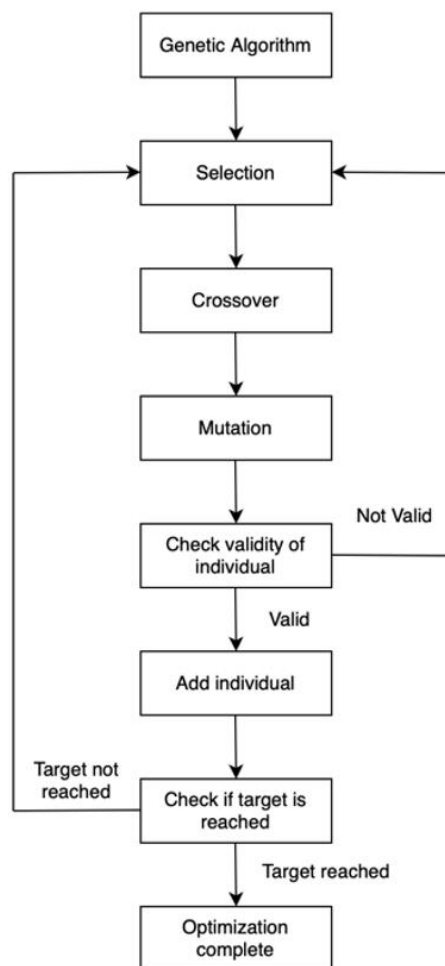


Figure 2. The process used during Genetic Algorithm [32]

The selection of Genetic Algorithms parameters, including population size, number of generations, mutation rate, and crossover rate, is essential for the optimization process's effectiveness. Here are a few factors to consider when selecting these parameters:

1. Population Size: An increased population size can augment genetic variety and elevate the likelihood of identifying optimal solutions. Nonetheless, it also escalates computational expenses. It is customary to initiate with a moderate population size (e.g., 50-100 persons) and modify based on initial findings [33].
2. Number of Generations: The quantity of generations dictates the duration of the algorithm's execution. This parameter may be adjusted according to the convergence behavior noted in the first runs. Should the population

exhibit considerable enhancement across generations, it could be advantageous to permit more generations [34].

3. Mutation and Crossover Rates: The mutation rate regulates the frequency of random alterations, while the crossover rate dictates the probability of merging solutions. The rates may be adjusted according to empirical findings, with typical initial values being a mutation rate of 0.01-0.1 and a crossover rate of 0.6-0.9 [35].

3. RESULTS AND DISCUSSION

3.1 Descriptive statistics

In this dataset, a total of 15 attributes have been identified as influential factors contributing to the occurrence of preeclampsia in mothers. Meanwhile, the target variable is the occurrence of preeclampsia which are an individual diagnosed with preeclampsia is referred to as a case, while an individual without preeclampsia is categorized as a control. Table 1 displays the descriptive statistics of the numerical attributes.

Table 1. Descriptive statistics of numeric attributes

Variable	Minimum	Maximum	Mean
Maternal age	17	42	26
Age at first pregnancy	14	38	22
Hemoglobin level	6.40	15.80	11.91

Based on the results presented in Table 1, specific observations can be made regarding the relationship between maternal age, hemoglobin level, and the risk of preeclampsia. Firstly, the data reveals that pregnant women under 20 years old or above 40 years old are more susceptible to experiencing preeclampsia. This suggests that extreme maternal age can contribute to increasing the risk of preeclampsia during pregnancy.

Additionally, the average hemoglobin level observed in the study is 11.91 grams per deciliter, which falls within the range of a normal hemoglobin level according to the WHO guidelines [36]. However, it is worth noting that vigilance is required for pregnant women whose hemoglobin levels dip below 10.9 grams per deciliter, as they may be at an increased risk of developing anemia.

These findings emphasize the importance of monitoring maternal age and hemoglobin levels during pregnancy to identify individuals more susceptible to preeclampsia or other related complications. Regular assessment and appropriate interventions, such as iron supplementation or medical supervision, can help mitigate the risks associated with maternal age and hemoglobin levels, thereby promoting better maternal and fetal health outcomes.

In addition to the numerical attributes discussed earlier, the categorical attributes are visually represented through the following Table 2 and graphs:

Table 2. Descriptive statistics of categorical attributes

Variable	Yes	No
Tobacco use	9	343
Alcohol use	25	327
Diabetes personal history	25	327
Diabetes family history	25	327
Hypertension personal history	25	327
Hypertension family history	55	297

Most respondents in this dataset, accounting for over 90%, reported abstaining from smoking and alcohol consumption. Additionally, a significant portion of the participants indicated non personal or family history of diabetes or hypertension. These findings suggest a low prevalence of these risk factors among the surveyed population.

The bar graph in Figure 3 depicts the distribution of two separate cohorts, labeled “Case” and “Control,” based on their various statuses. The “Control” cohort comprises roughly 264 individuals, whereas the “Case” cohort includes about 88 individuals. This observation reveals that the dataset has more persons designated as “Control” relative to those identified as “Case,” underscoring a significant disparity between the two groups. In the context of analysis, especially in machine learning or statistical modeling, such an imbalance may require rectification since it could negatively impact the performance of the utilized models.

The dataset demonstrates a considerable disparity between the control and case classifications of preeclampsia. To resolve this issue, we employed the Synthetic Minority Over-sampling Technique (SMOTE) to attain balance among the categories by generating synthetic instances for the underrepresented class. SMOTE improves the efficacy of machine learning models by generating a more balanced dataset. SMOTE enhances algorithms’ understanding of the minority class by generating synthetic examples, enhancing accuracy, precision, and recall. Zhou et al. [37] showed that using SMOTE in their dataset enhanced the predictive performance of models, including Logistic Regression, Random Forest, and Support Vector Machines. Although SMOTE produces synthetic samples, it may inject noise into the dataset if the minority class cases are inadequately represented. This noise may result in diminished model performance, mainly if the synthetic samples are produced in regions of the feature space that fail to accurately depict the genuine distribution of the minority class [38].

We partitioned the dataset into training and testing subsets, designating 80% for training and 20% for evaluation. The training subset was used to construct the model, while the testing subset was applied to assess its effectiveness.

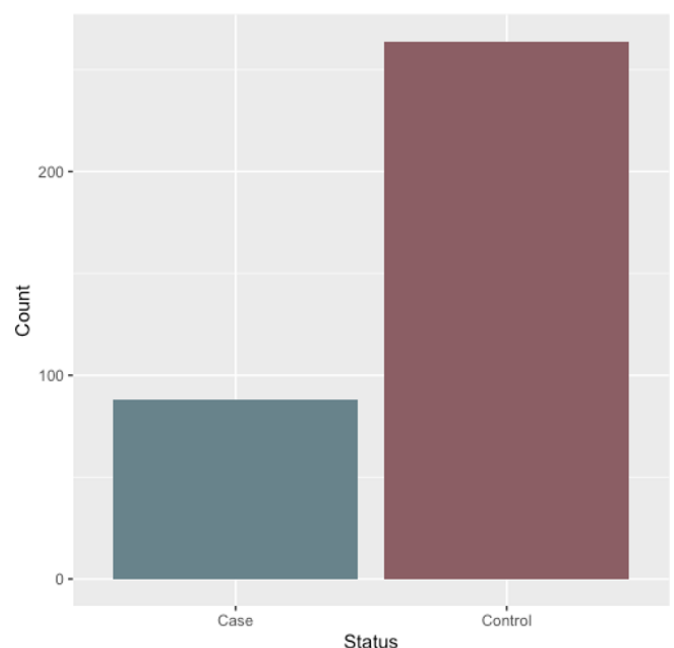


Figure 3. Status of preeclampsia

3.2 Predictive modeling

We employ the Random Forest model and want to adjust the hyperparameters *n*tree and *m*try. Since *m*try and *n*tree are discrete value hyperparameters, the optimization procedure employs the binary encoding. Here, we have set the *m*try range to 1 to 8 and the *n*tree range to 1 to 512.

To optimize these parameters, we use a Genetic Algorithm with several conditions that we have set, which are shown in the Table 3.

Table 3. Conditions setting in Genetic Algorithm

Conditions	Value
Type	binary
Fitness	<i>fit_rf</i>
Nbits	14
Population size	50
Iterations	30

The Genetic Algorithm initiates by generating an initial population of candidate solutions, each representing a distinct set of hyperparameters for the Random Forest model. Each candidate solution is assessed using a fitness function that generally quantifies the efficacy of the Random Forest model using a designated metric, such as accuracy. The fitness score reflects the model’s performance using the specified hyperparameters. The evaluation process is conducted for every population member [39]. Upon assessing the viability of each prospective solution, the Genetic Algorithm selects individuals to serve as parents for the subsequent generation. The chosen parent solutions undergo crossover, wherein pairs of parents exchange segments of their hyperparameter configurations to generate offspring. The Genetic Algorithm implements mutation in particular offspring to preserve genetic variety within the population and avert early convergence. The evaluation, selection, crossover, and mutation procedures are reiterated for a certain number of generations or until a termination criterion is satisfied. This iterative procedure enables the Genetic Algorithm to converge on optimal hyperparameter configurations for the Random Forest model [40].

We use binary type in this Genetic Algorithm since the possible representation of decision variables is only two—cases of preeclampsia and control. We create a fitness function in *fit_rf* that represents the best parameter search criteria in a Random Forest model. By multiplying and adding the maximum value of each hyperparameter to the total number of hyperparameters, we can determine the number of bits we require, and in this case is *Nbits* = 14. These values indicate how many bits should be utilized in binary-encoded optimizations.

Figure 4 provides the best fitness value results for 30 iterations or generations.

From Figure 4 and Figure 5, it can be seen that in the initial generation with *m*try 1 and *n*tree 35, the accuracy (fitness value) only reached around 0.65. Then it continues to increase until it reaches the highest value in the 30th generation with a *m*try value of 4 and *n*tree of 9 with an accuracy of 0.90. This result was obtained by applying the number of best-fitness individuals to survive at each generation to 2 with crossover probability = 0.8 and mutation probability = 0.1.

Next, using the best Random Forest parameters obtained from the Genetic Algorithm process, preeclampsia status classification modeling was carried out with an accuracy of

75% with the Random Forest formed as follows.

Figure 6 depicts the optimal decision tree constructed by the Random Forest algorithm. The initial bifurcation is influenced by maternal education, underscoring its importance as a pivotal variable. Subsequently, the left node undergoes further subdivision predicated upon a personal history of hypertension, whereas the right node is forked according to the age of pregnancy. Additional divisions are executed utilizing other variables, thereby progressively enhancing the precision of the predictions. Ultimately, the tree culminates in terminal nodes that categorize the preeclampsia status, offering valuable insights into the predictive methodology.

According to Figure 7, both *MeanDecreaseAccuracy* and *MeanDecreaseGini* help us understand the significance of individual features in a Random Forest model. *MeanDecreaseAccuracy* provides insights into how characteristics affect overall model accuracy, while *MeanDecreaseGini* focuses on their ability to reduce impurity and improve class separation in the decision trees. The higher the value of *MeanDecreaseAccuracy* or *MeanDecreaseGini* score, the higher the importance of the variable in the model. In the Figure 7, age at first pregnancy, maternal age, hemoglobin level, hypertension, personal history, and ethnicity are five essential variables.

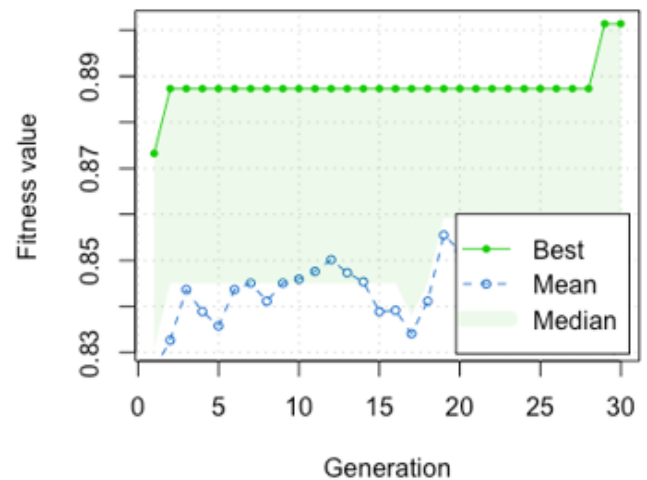


Figure 4. The fitness value of each generation

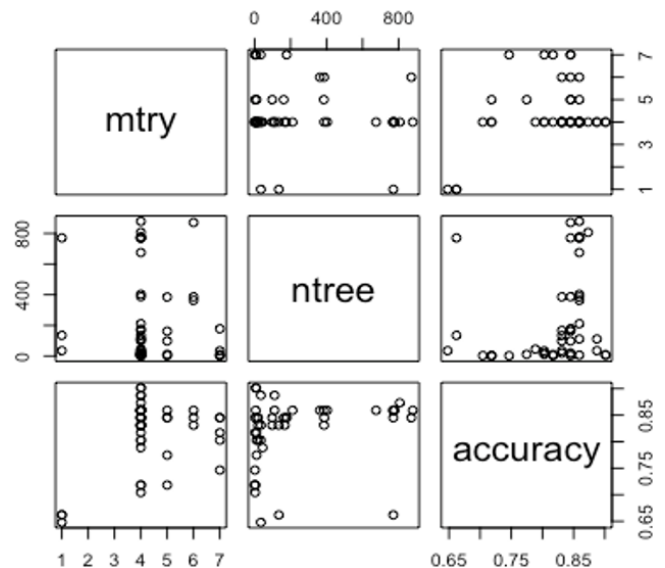


Figure 5. The accuracy of each parameter

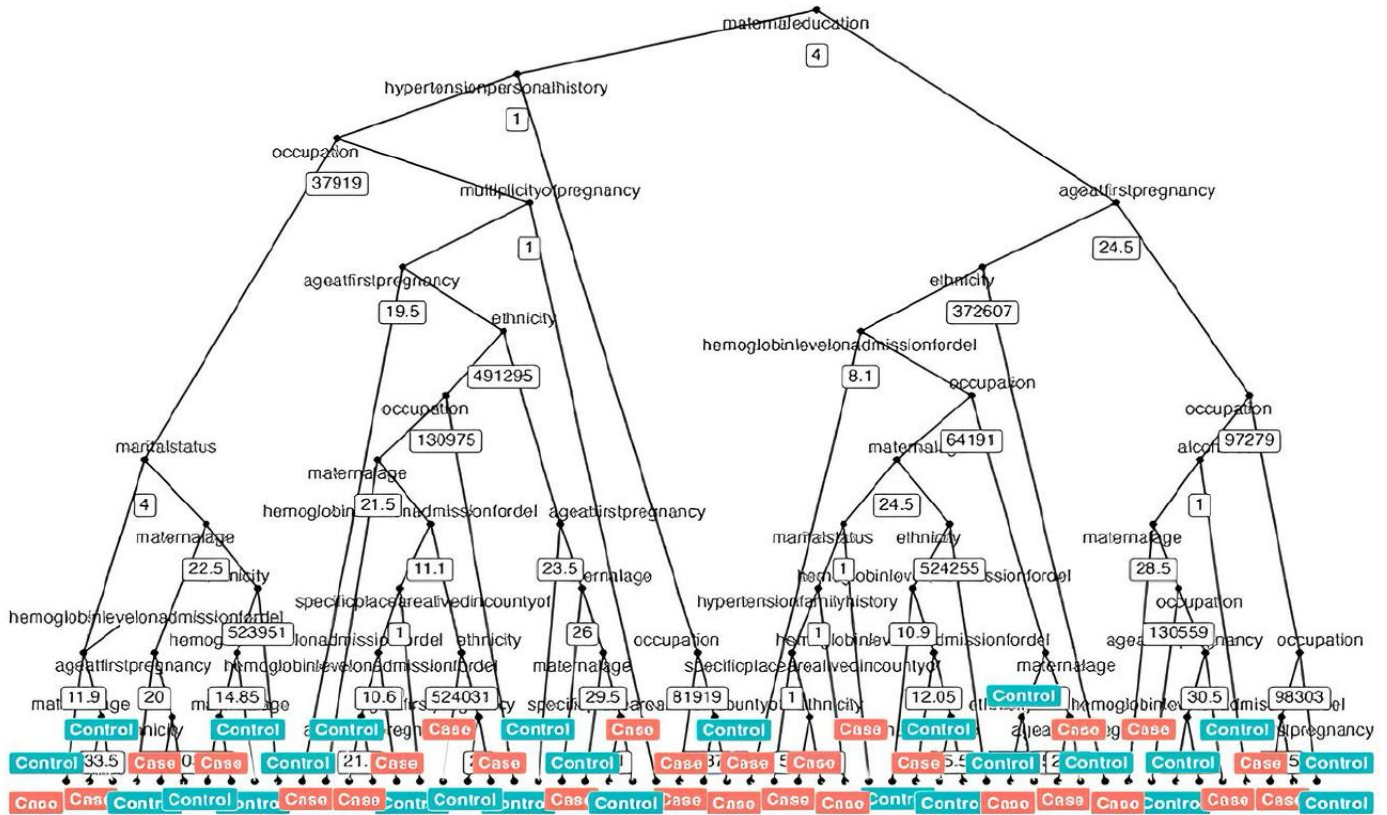


Figure 6. Random Forest

Variable Importance Plot

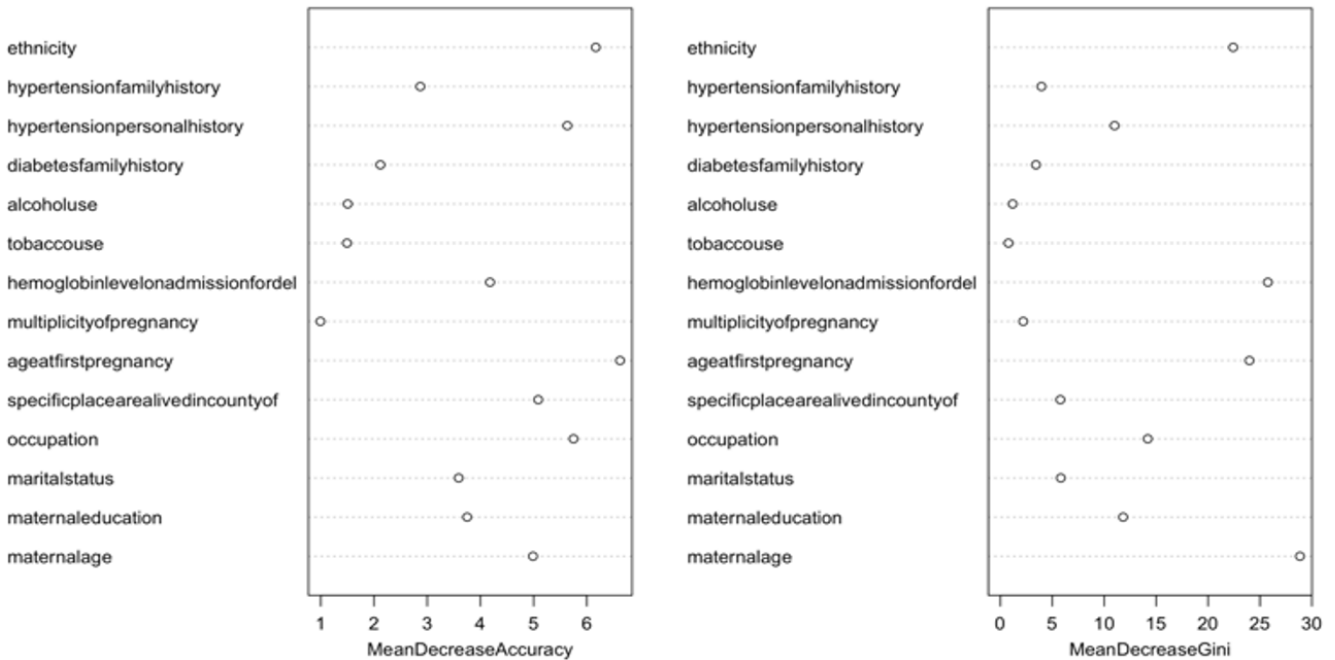


Figure 7. Variable importance

3.3 The applications system

Using the Shiny interface, we created a desktop application to run the Random Forest with Genetic Algorithm on a personal computer through RStudio. Users can access the application through this address

<https://pstat.shinyapps.io/deploytesting/>. On the main page, users are given the option that they can input data partially or entirely. We also provide a user guide, making it easier for users to use this application (Figures 8-9).

For instance, suppose there is a patient with the following history: Using the patient's history, as in Table 4, the

application will predict that the patient is at risk of preeclampsia (case of preeclampsia). In the same way, the application can predict the occurrence of preeclampsia at once for several patients, as shown in Figure 10.

We intend to deploy our methodology in partnership with a hospital in our region. This will enable us to assess its efficacy against existing screening methods and analyze its influence on clinical results. This real-world application will yield significant insights into how the model can be enhanced to better align with practical requirements in clinical environments. This implementation phase may reveal essential enhancements to improve the model's usability and integration into standard healthcare procedures, ultimately seeking to increase early diagnosis and intervention outcomes for preeclampsia.

Table 4. Example 1

Variable	Value	Variable	Value
Maternal age	30	Hypertension family history	Yes
Age at first pregnancy	28	Multiplicity of pregnancy	Twin
Hemoglobin level	16	Maternal education	Primary education
Tobacco use	Yes	Marital status	Married
Alcohol use	No	Occupation	Housewife
Diabetes family history	No	Area lived	Rural
Hypertension personal history	Yes	Ethnicity	Taita

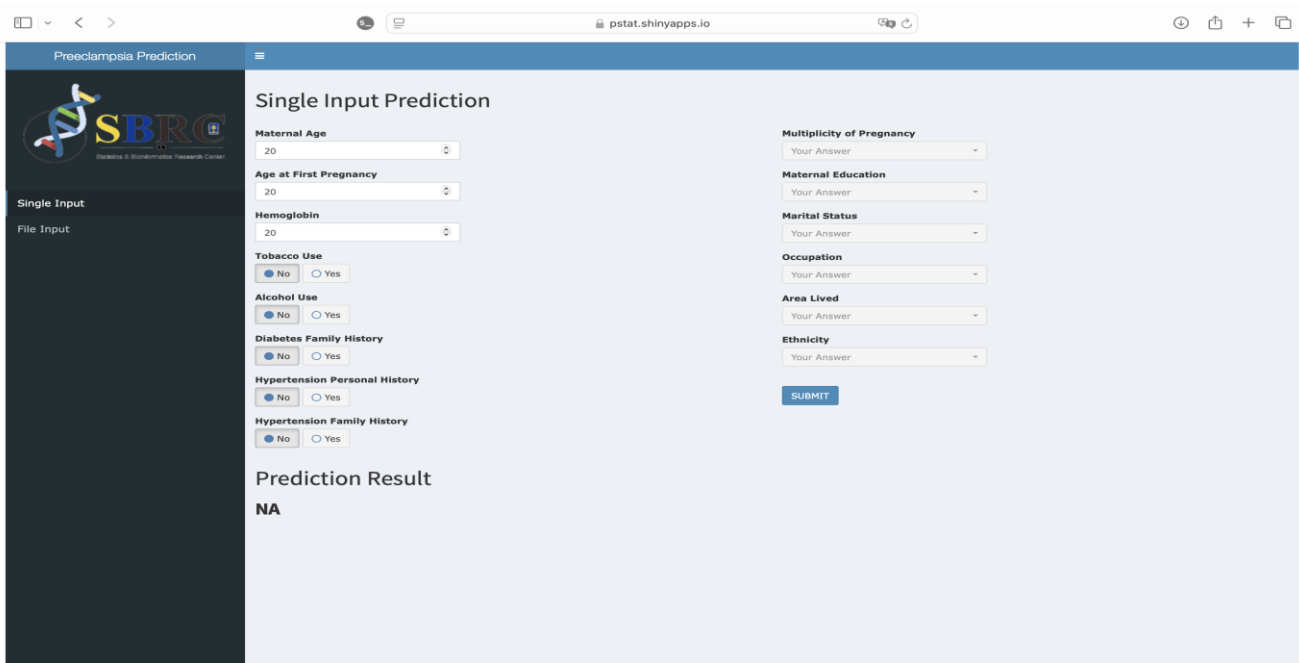


Figure 8. Main page of the application

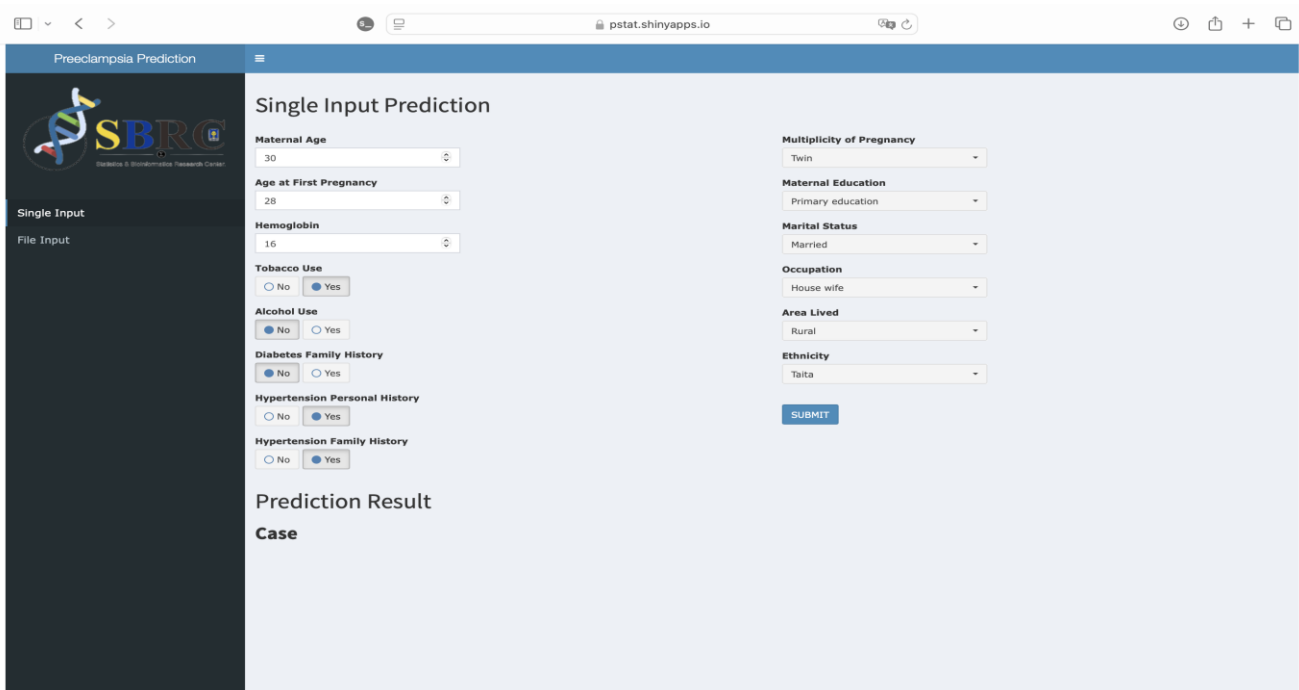


Figure 9. Example of partial input

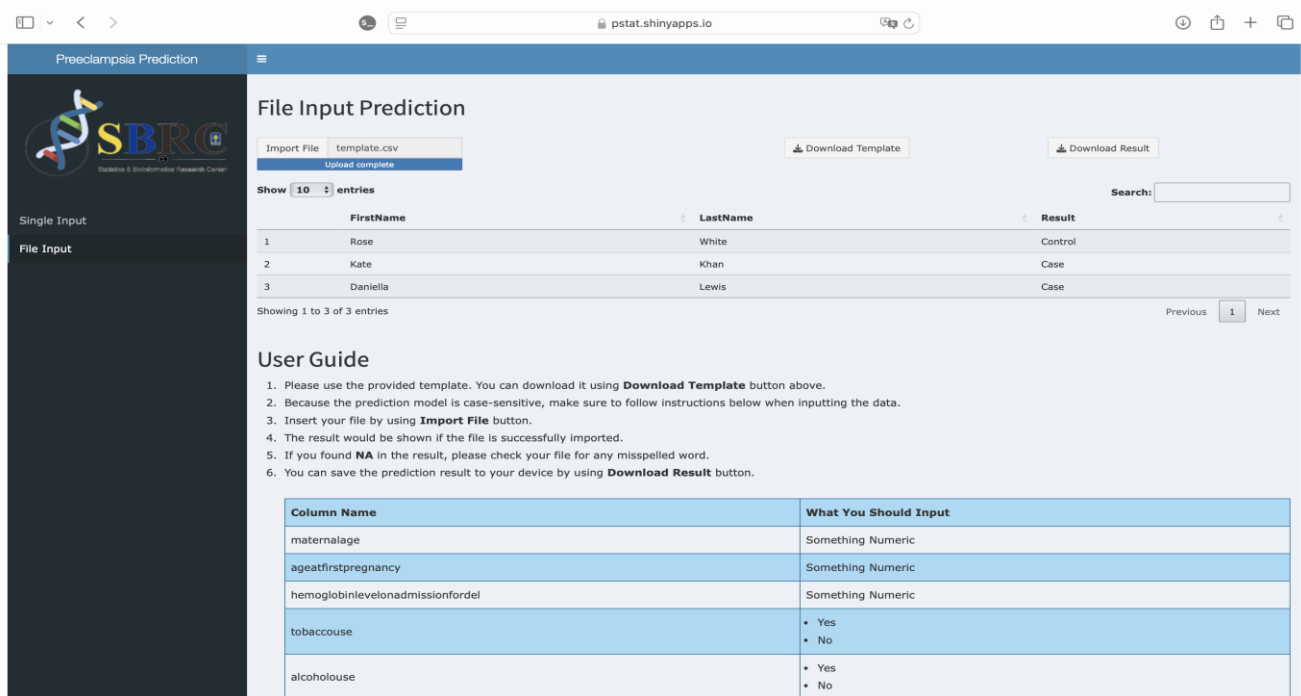


Figure 10. Example of entire input

4. CONCLUSIONS

This research highlights an important aspect that contribute to improving the Random Forest model: parameter optimization using Genetic Algorithm. The application of Genetic Algorithm for parameter optimization in the Random Forest model has shown promising results. Genetic Algorithms offer a robust and efficient approach to searching the parameter space, identifying the optimal combination of parameters that maximize the model's performance. The Random Forest model can be significantly improved in terms of predictive power and generalization capability by fine-tuning the parameters.

Furthermore, the development and utilization of an R-Shiny App for early detection of preeclampsia offer several significant benefits in maternal healthcare. This interactive and user-friendly application provides a streamlined and accessible platform for healthcare professionals, making identifying and assessing preeclampsia risk more efficient and accurate. With its intuitive interface, medical practitioners can input patient data and receive real-time predictions, enabling earlier detection and intervention, which is crucial for maternal and fetal well-being.

This research emphasizes accuracy as a performance indicator; however, we acknowledge that future endeavors could enhance the model and its application by integrating additional metrics to yield a more nuanced comprehension of the model's efficacy in identifying preeclampsia risk.

Future studies should examine and contrast the model's performance across various demographic groups to mitigate any biases and improve its applicability to diverse populations. This approach would facilitate an assessment of the model's generalizability and guarantee uniform performance across diverse subpopulations. Incorporating these concerns in future studies will bolster the rigor of this research and augment the findings' contribution to the field.

ACKNOWLEDGMENT

The authors would like to acknowledge the Directorate of Research and Community Services, Universitas Islam Indonesia for their financial and others valuable supports in this research.

REFERENCES

- [1] Jhee, J.H., Lee, S., Park, Y., Lee, S.E., Kim, Y.A., Kang, S.W., Kwon, J.A., Park, J.T. (2019). Prediction model development of late-onset preeclampsia using machine learning-based methods. *PLoS One*, 14(8): e0221202. <https://doi.org/10.1371/journal.pone.0221202>
- [2] Tahir, M., Badriyah, T., Syarif, I. (2018). Neural networks algorithm to inquire previous preeclampsia factors in women with chronic hypertension during pregnancy in childbirth process. In 2018 International Electronics Symposium on Knowledge Creation and Intelligent Computing (IES-KCIC), Bali, Indonesia, pp. 51-55. <https://doi.org/10.1109/KCIC.2018.8628588>
- [3] Moreira, M.W., Rodrigues, J.J., Marcondes, G.A., Neto, A.J.V., Kumar, N., Diez, I.D.L.T. (2018, May). A preterm birth risk prediction system for mobile health applications based on the support vector machine algorithm. In 2018 IEEE International Conference on Communications (ICC), Kansas City, USA, pp. 1-5. <https://doi.org/10.1109/ICC.2018.8422616>
- [4] C Ramdhani, Y., Maulidia, D., Setiadi, A., Alamsyah, D.P. (2022). Feature weighting optimization: genetic algorithms and random forest for classification of pregnant potential risk. In 2022 International Conference on Information Technology Research and Innovation (ICITRI), Jakarta, Indonesia, pp. 95-100. <https://doi.org/10.1109/ICITRI56423.2022.9970206>
- [5] Gollapalli, M., Rahman, A., Youldash, M., Alomar, D., Alismail, S., Khawaher, F., Alkhadair, A., Aljubran, F.,

- Alzannan, R., Alkhulaifi, D., Mahmud, M. (2023). Machine learning approach to users' age prediction: A telecom company case study in Saudi Arabia. *Mathematical Modelling of Engineering Problems*, 10(5): 1619–1629. <https://doi.org/10.18280/mmep.100512>
- [6] Berk, R.A. (2016). Random forests. In *Statistical Learning from a Regression Perspective*, Springer Cham, Singapore, pp. 205-258. https://doi.org/10.1007/978-3-319-44048-4_5
- [7] Srinivas, M., Patnaik, L.M. (1994). Genetic algorithms: A survey. *Computer*, 27(6): 17-26. <https://doi.org/10.1109/2.294849>
- [8] Dinakaran, S., Thangaiah, P.R.J. (2014). Comparative analysis of filter-wrapper approach for random forest performance on multivariate data. In *2014 International Conference on Intelligent Computing Applications*, Coimbatore, India, pp. 174-178. <https://doi.org/10.1109/ICICA.2014.45>
- [9] Melinte-Popescu, A.S., Vasilache, I.A., Socolov, D., Melinte-Popescu, M. (2023). Predictive performance of machine learning-based methods for the prediction of preeclampsia—A prospective study. *Journal of Clinical Medicine*, 12(2): 418. <https://doi.org/10.3390/jcm12020418>
- [10] Sufriyana, H., Husnayain, A., Chen, Y.L., Kuo, C.Y., Singh, O., Yeh, T.Y., Wu, Y.W., Su, E.C.Y. (2020). Comparison of multivariable logistic regression and other machine learning algorithms for prognostic prediction studies in pregnancy care: Systematic review and meta-analysis. *JMIR Medical Informatics*, 8(11): e16503. <https://doi.org/10.2196/16503>
- [11] Loftness, B.C., Bernstein, I., McBride, C.A., Cheney, N., McGinnis, E.W., McGinnis, R.S. (2023). Preterm preeclampsia risk modelling: Examining hemodynamic, biochemical, and biophysical markers prior to pregnancy. In *2023 45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, Sydney, Australia, pp. 1-4. <https://doi.org/10.1109/EMBC40787.2023.10340404>
- [12] Zheng, Y.H., Fang, Z.J., Wu, X.Z., Zhang, H.L., Sun, P.M. (2023). Identification of F13A1 and SCCPDH as potential diagnostic markers for preeclampsia. *Research Square*. <https://doi.org/10.21203/rs.3.rs-3407760/v1>
- [13] Thrash, A., Tang, J.D., DeOrnellis, M., Peterson, D.G., Warburton, M.L. (2020). PAST: The pathway association studies tool to infer biological meaning from GWAS datasets. *Plants*, 9(1): 58. <https://doi.org/10.3390/plants9010058>
- [14] Simkin, J., Dummer, T.J., Erickson, A.C., Otterstatter, M.C., Woods, R.R., Ogilvie, G. (2022). Small area disease mapping of cancer incidence in British Columbia using Bayesian spatial models and the smallareamapp R Package. *Frontiers in Oncology*, 12: 833265. <https://doi.org/10.3389/fonc.2022.833265>
- [15] Wang, Y.A., Rhemtulla, M. (2021). Power analysis for parameter estimation in structural equation modeling: A discussion and tutorial. *Advances in Methods and Practices in Psychological Science*, 4(1). <https://doi.org/10.1177/2515245920918253>
- [16] Heiss-Czedik, D. (1997). An introduction to genetic algorithms. *Artificial Life*, 3(1): 63-65. <https://doi.org/10.1162/artl.1997.3.1.63>
- [17] Nijhout, F. (1997). An introduction to genetic algorithms. *Complex*, 2(5): 39-40. [https://doi.org/10.1002/\(SICI\)1099-0526\(199705/06\)2:5%3C39::AID-CPLX8%3E3.0.CO;2-L](https://doi.org/10.1002/(SICI)1099-0526(199705/06)2:5%3C39::AID-CPLX8%3E3.0.CO;2-L)
- [18] Goldberg, D.E. (1989). *Genetic Algorithm in Search, Optimization and Machine Learning*, Addison. Wesley Publishing Company.
- [19] Logan, G.G., Njoroge, P.K., Nyabola, L.O., Mweu, M.M. (2020). Determinants of preeclampsia and eclampsia among women delivering in county hospitals in Nairobi, Kenya. *F1000Research*, 9: 192. <https://doi.org/10.12688/f1000research.21684.1>
- [20] Breiman, L. (2001). Random forests. *Machine Learning*, 45(1): 5-32. <https://doi.org/10.1023/A:1010933404324>
- [21] Nguyen, H.T.T., Chen, L.H., Saravanarajan, V.S., Pham, H.Q. (2021). Using XG boost and random forest classifier algorithms to predict student behavior. In *2021 Emerging Trends in Industry 4.0*, Raigarh, India, pp. 1-5. <https://doi.org/10.1109/ETI4.051663.2021.9619217>
- [22] Gollapalli, M., Alqahtani, T.A., Alhamed, D.H., Alnassar, M.R., Alajmi, A.M., Alali, Y.H., Abdulkader, M.M., Saadeldin, A. (2023). Intelligent modelling techniques for predicting used cars prices in Saudi Arabia. *Mathematical Modelling of Engineering Problems*, 10(1): 139-148. <https://doi.org/10.18280/mmep.100115>
- [23] Jaiswal, J.K., Samikannu, R. (2017). Application of random forest algorithm on feature subset selection and classification and regression. In *2017 world congress on computing and communication technologies (WCCCT)*, Tiruchirappalli, India, pp. 65-68. <https://doi.org/10.1109/WCCCT.2016.25>
- [24] Binarwati, L., Mukhlash, I., Soetrisno, S. (2017). Implementasi algoritma genetika untuk optimalisasi random forest dalam proses klasifikasi penerimaan tenaga kerja baru: Studi kasus PT. XYZ. *Jurnal Sains dan Seni ITS*, 6(2): A78-A83. <https://doi.org/10.12962/j23373520.v6i2.26887>
- [25] Cutler, A., Cutler, D.R., Stevens, J.R. (2012). Random forests. *Ensemble Machine Learning: Methods and Applications*. Springer, USA, pp. 157-175. https://doi.org/10.1007/978-1-4419-9326-7_5
- [26] Mitchell, M. (1995). Genetic algorithms: An overview. *Complex*, 1(1): 31-39. <https://doi.org/10.1002/cplx.6130010108>
- [27] Muniasamy, K., Venugopal, P., Pakkirisamy, G. (2023). Genetic algorithm-driven optimization of scheduling and preventive measures in parallel machines. *Mathematical Modelling of Engineering Problems*, 10(5): 1811-1816. <https://doi.org/10.18280/mmep.100533>
- [28] Shi, T., He, G., Mu, Y.L. (2019). Random forest algorithm based on genetic algorithm optimization for property-related crime prediction. In *2019 International Conference on Computer, Network, Communication and Information Systems (CNCI 2019)*, Qingdao, China, pp. 526-531. <https://doi.org/10.2991/cnci-19.2019.73>
- [29] El-Shafiey, M.G., Hagag, A., El-Dahshan, E.S.A., Ismail, M.A. (2022). A hybrid GA and PSO optimized approach for heart-disease prediction based on random forest. *Multimedia Tools and Applications*, 81(13): 18155-18179. <https://doi.org/10.1007/s11042-022-12425-x>
- [30] Di Francescomarino, C., Dumas, M., Federici, M., Ghidini, C., Maggi, F.M., Rizzi, W., Simonetto, L.

- (2018). Genetic algorithms for hyperparameter optimization in predictive business process monitoring. *Information Systems*, 74: 67-83. <https://doi.org/10.1016/j.is.2018.01.003>
- [31] Li, C., Jiang, J.Z., Zhao, Y.Q., Li, R.G., Wang, E.D., Zhang, X., Zhao, K. (2022). Genetic algorithm based hyper-parameters optimization for transfer convolutional neural network. In *International Conference on Advanced Algorithms and Neural Networks (AANN 2022)*, Zhuhai, China, pp. 232-241. <https://doi.org/10.1117/12.2637170>
- [32] Ganapathy, K. (2020). A study of genetic algorithms for hyperparameter optimization of neural networks in machine translation. arXiv: 2009.08928. <https://doi.org/10.48550/arXiv.2009.08928>
- [33] Chen, J., Okle, R.A.N. (2024). A study on the application of response surface methodology and Bayesian optimization in parameter tuning of genetic algorithms. In *Fourth International Conference on Applied Mathematics, Modelling, and Intelligent Computing (CAMMIC 2024)*, Kaifeng, China, pp. 143-150. <https://doi.org/10.1117/12.3036921>
- [34] Liu, G.Y., Hou, Y.B., Luo, Y., Li, D. (2017). Genetic algorithm's application for optimization of PID parameters in dynamic positioning vessel. In *MATEC Web of Conferences*, Chengdu, China, pp. 00153. <https://doi.org/10.1051/mateconf/201713900153>
- [35] Syukron, A., Subekti, A. (2018). Penerapan metode random over-under sampling dan random forest untuk klasifikasi penilaian kredit. *Jurnal Informatika*, 5(2): 175-185. <https://doi.org/10.31294/ji.v5i2.4158>
- [36] World Health Organization. (2024). Anaemia, World Health Organization. https://www.who.int/health-topics/anaemia#tab=tab_1, accessed on Jun. 20, 2024.
- [37] Zhou, R.S., Yin, W.H., Li, W.J., Wang, Y.C., Lu, J., Li, Z., Hu, X.X. (2022). Prediction model for infectious disease health literacy based on synthetic minority oversampling technique algorithm. *Computational and Mathematical Methods in Medicine*, 2022(1): 8498159. <https://doi.org/10.1155/2022/8498159>
- [38] Herawati, B.C., Hairani, H., Guterres, J.X. (2024). SMOTE variants and random forest method: A comprehensive approach to breast cancer classification. *International Journal of Engineering Continuity*, 3(1): 12-23. <https://doi.org/10.58291/ijec.v3i1.147>
- [39] Probst, P., Wright, M.N., Boulesteix, A.L. (2019). *Hyperparameters and tuning strategies for random forest*. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 9(3): e1301. <https://doi.org/10.1002/widm.1301>
- [40] Raji, I.D., Bello-Salau, H., Umoh, I.J., Onumanyi, A.J., Adegboye, M.A., Salawudeen, A.T. (2022). Simple deterministic selection-based genetic algorithm for hyperparameter tuning of machine learning models. *Applied Sciences*, 12(3): 1186. <https://doi.org/10.3390/app12031186>