JURNAL INFOTEL
Informatics - Telecommunication - Electronics
Website: https://ejournal.ittelkom-pwt.ac.id/index.php/infotel
ISSN: 2085-3688; e-ISSN: 2460-0997

Improving malaria prediction with random forest and robust scaler: An integrated approach for enhanced accuracy

Azka Khoirunnisa[1,*], Nur Ghaniaviyanto Ramadhan[2]
[1]Department of Data Science, Telkom University
[2]Department of Software Engineering, Institut Teknologi Telkom Purwokerto
[1]Jl. Telekomunikasi, Kabupaten Bandung 40257, Indonesia
[2]Jl. D. I. Panjaitan, No. 128, Purwokerto 53147, Indonesia
[*]Corresponding email: khoirunnisaazka@telkomuniversity.ac.id

Received 16 October 2023, Revised 28 October 2023, Accepted 2 November 2023

Abstract — Mosquito bites are the primary transmission method for malaria, a prevalent and significant health concern worldwide. In the context of malaria incidence, Indonesia is the second most affected country after India. According to the Ministry of Health's report, Papua Province reported 216,380 malaria cases in 2019. Additionally, East Nusa Tenggara and West Papua said 12,909 and 7,029 points, respectively, reflecting the substantial national burden of this disease. Predicting malaria occurrence based on symptomatic presentation is a crucial preventive strategy. Machine learning models offer a promising approach to malaria prediction. This study focused on malaria detection by using patient data from Nigeria. This research proposes a detection system utilizing the Random Forests method, employing Robust Scaler for effective normalization, and integrating K-fold cross-validation to enhance model robustness. Various experiments were conducted by systematically varying K values and the number of decision trees to ascertain the most effective hyper-parameters yielding the highest accuracy. The findings indicate that the optimal accuracy of 82 % is achieved at a k value of 20, showing comparable accuracies across different decision tree quantities, underlining the robustness of the employed method. This research significantly advances malaria detection strategies, offering valuable insights into the effective deployment of machine learning in health-care decision-making.

Keywords – classification, malaria, random forests, robust scaler, k-cross validation

I. Introduction

Malaria is a disease caused by mosquito bites [1]. Malaria in Indonesia ranks second highest after India [2]. Based on the Ministry of Health's report, provinces in Indonesia with high malaria cases in 2019 include Papua province, with 216,380 cases; East Nusa Tenggara, with 12,909 cases, and West Papua with 7,029 cases. The high number of malaria cases in Indonesia, especially in the eastern region, is important for prevention. One-way Preventive measures can be done by predicting whether someone is affected by malaria based on the symptoms experienced.

Malaria prediction can be done using machine learning models [3]–[5]. The study [6] aimed to predict malaria based on a patient's clinical information by comparing six machine learning models, including random forest (RF) and multi-layer perceptron (MLP). The study diagnosed malaria according to symptoms and analyzed important features of inpatient data using the best predictive model RF [7]. Some features used include fever, headache, age, and general body malaise. The research also applies the K-fold cross-validation technique. the study predicted malaria based on the symptoms a person was experiencing [8]. Several machine learning models, including RF, support vector machine (SVM), and artificial neural network (ANN), were utilized for this prediction. The study also handles data preprocessing tasks such as cleansing null values, normalization, and balancing the data count.

The research [9] detected malaria using convolutional neural network (CNN) deep learning models. The study [10]used machine learning models in depth to reduce the impact and predict malaria. The model used is long short memory term (LSTM). The research [11] diagnosed malaria based on Symptoms data. The model

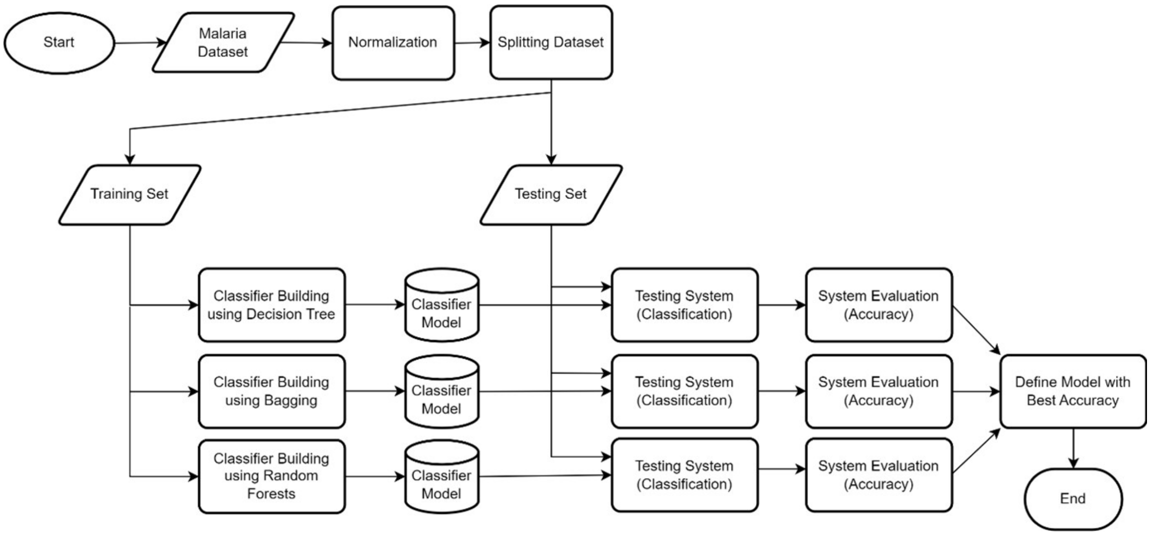
Start
Malaria Dataset
Normalization
Splitting Dataset
Training Set
Testing Set
Classifier Building using Decision Tree
Classifier Building using Bagging
Classifier Building using Random Forests
Classifier Model
Classifier Model
Classifier Model
Testing System (Classification)
Testing System (Classification)
Testing System (Classification)
System Evaluation (Accuracy)
System Evaluation (Accuracy)
System Evaluation (Accuracy)
Define Model with Best Accuracy
End

Fig. 1. System design.

applied is the decision tree (DT). Feature analysis of symptoms resulted in an accuracy of 77 %. The other study [12] classified symptoms for malaria prediction. The three models used are neural network (NN), SVM, and regression logistics (RL). The study [13] classified malaria data using the classification and regression tree (CART) and Naive Bayes (NB) models. Preprocessing techniques include data normalization using min-max scaling and feature correlation analysis.

The research [14] identified malaria using a CNN algorithm. The research involves feature extraction and feature selection on the dataset used. The research [15] utilized a hybrid model of SVM and adaptive boosting (AdaBoost) for malaria diagnosis. Data preprocessing techniques include eliminating redundant values and feature extraction using chi-square. The research [16] developed an adaptive genetic algorithm model for malaria data classification. The recursive features elimination (RFE) model is used to select relevant features in the data. This study took advantage of opportunities in previous research gaps, such as applying robust normalization techniques and feature correlation analysis.

Based on an explanation of the issues with earlier studies, this study has the main contributions, including normalizing the dataset using a robust scaler, investigating correlation relationships among data features, and predicting malaria using the RF model. In this paper, there are several sections. Section I serves as an introduction, discussing several issues in previous research. Section II is the related works, which discusses research closely related to the proposed study. Section III presents the proposed research method along with the research diagram. Section IV discusses the results and analysis of the conducted experiments. Finally, Section V serves as the conclusion of this research.

II. Research Method

In this section, we explain the dataset, feature correlations, normalization techniques, and predictive models used in this study. Fig. 1 shows the flowchart that illustrates the key steps and processes involved in our research methodology.

A. Dataset

The dataset used in this research is the data of malaria patients in Nigeria [17]. This dataset consists of 18 features related to malaria patients, along with one class feature. A representation of the dataset employed is displayed in Table 1.

Table 1. Overview of the Datasets

No	Features	Features Type
1	Age	Numerical
2	Sex	Binary (0,1)
3	Fever	Binary (0,1)
4	Cold	Binary (0,1)
5	Rigor	Binary (0,1)
6	Fatigue	Binary (0,1)
7	Headache	Binary (0,1)
8	Bitter_tounge	Binary (0,1)
9	Vomiting	Binary (0,1)
10	Diarrhea	Binary (0,1)
11	Convulsion	Binary (0,1)
12	Anemia	Binary (0,1)
13	Jaundice	Binary (0,1)
14	Cocacola_urine	Binary (0,1)
15	Hypoglycemia	Binary (0,1)
16	Prostraction	Binary (0,1)
17	Hyperpyxeria	Binary (0,1)
18	Severe_melaria	Binary (0,1)

B. Features Correlation

Feature correlation in classification refers to the relationship or dependency between various features or variables used in a classification model. These features can be attributes or characteristics used to describe the data to be classified. Feature correlation measures the extent to which these features are related to each other or how one feature influences another [18], [19]. In this paper, all the features in the dataset are used because these features are important for determining malaria.

C. Normalization

Normalization is changing the values of features or attributes in a dataset so that they have a uniform scale or range. Normalization is employed to achieve the highest possible accuracy in classification. This process serves to eliminate features that contain excessive noise or bear little relevance to the class [20]–[22]. The normalization technique used in this research is Robust Scaler. This technique scales features by dealing with outliers in the data. The formula used in the robust scaler is shown in (1) [23].

$$x_{scaled} = \frac{(x - Q2\,(x))}{(Q3\,(x) - Q1\,(x))} \qquad (1)$$

where x_{scaled} is scaled and normalized value of the feature x, x is original data point of the feature, $Q2(x)$ is median (the second quartile) of the feature x, $Q1(x)$ is first quartile of the feature x, and $Q3(x)$ is third quartile of the feature x.

D. Decision Trees

Decision trees are one of the most used classification methods in machine learning and data analysis. They are models that separate and classify data based on hierarchical decision rules. Decision trees are easy to interpret and can be used in various of applications. Decision trees are built by dividing the dataset into increasingly smaller subsets based on existing attributes. This process continues until it reaches the point where all the data in each branch of the tree belongs to the same class or group [24].

Decision trees have the advantage of being easy to interpret, allowing a good understanding of how the model makes decisions. Additionally, they can handle categorical and numeric data, as well as being able to handle feature interactions. However, the disadvantage lies in the tendency towards over-fitting, especially when the tree grows very complex, and the possibility of creating too large trees. Selecting the most informative first attribute can also affect model performance, and class imbalance in the data can produce a tree that is biased towards the majority class. Therefore, pruning or the use of ensemble methods is often necessary to overcome these shortcomings and improve classification accuracy [25].

Furthermore, the decision trees algorithm application process generally involves the following steps:

1) Initialization: Start with the entire training dataset.
2) Select the best feature for the root node: a.
    a) Calculate the impurity or information gain (IG) for each feature in the dataset using (2).

    $$IG = E\,(S) - \sum_{v} \frac{|S_v|}{|S|} E\,(S_v) \qquad (2)$$

    where E is the Entropy.
    b) Choose the feature that results in the highest information gain or the lowest impurity as the root node of the tree.
3) Create a decision tree node associated with the selected feature.
4) Data split: Split the dataset based on the split attribute values selected in the previous step. Each attribute value will produce one branch in the tree.
5) Recursion: Continue steps 2 and 3 for each branch (subset) of the resulting dataset. Repeat these steps until the stop condition is met, such as:
    - All data on a branch belongs to the same class
    - The depth of the tree reaches a predetermined limit
    - And there are no attributes left for separation.
6) Pruning: After building the tree, you can prune insignificant branches to avoid over-fitting. Pruning can be done using various methods, such as tree depth reduction or pruning based on metric values.
7) Majority class determination: When reaching a leaf of the tree, or if there are no split attributes remaining, determine the majority class in that subset of data as the prediction label for that leaf.
8) Completed model: The decision tree model is completed once this process is completed.

After the decision tree model has been built, the model can be used to classify new data by following the rules in the tree. This process involves traversing the tree from root to leaf according to the attribute values of the data to be predicted.

E. Bagging (Bootstrap Aggregating)

The bagging (bootstrap aggregating) classification method is an ensemble technique in machine learning that utilizes random repetition (bootstrap) to create several subsets of the training data. Each subset is used to train the same or similar classification models, such as Decision Trees. Predictions from each model are combined, often using of majority voting, to produce a final prediction. Bagging aims to reduce over-fitting and model variance by leveraging insights from multiple models, ultimately improving classification accuracy and stability [26].

The advantage of the Bagging classification method is that it is effective in reducing over-fitting and model variance, resulting in more stable and accurate predictions. By combining insights from multiple models trained on different subsets of data, Bagging improves a model's ability to generalize data it has never seen before. However, its drawback lies in the increased in computational complexity as it involves training several similar models, which can be time and resource-consuming. Additionally, Bagging may be less effective if used on very small datasets or on very weak base models, as there may not be enough variation in the resulting models [27].

The Bagging algorithm application process generally involves the following steps [28]:

1) Initialization: Start with the original training dataset.
2) Create multiple subsets: Perform a bootstrap process to create multiple random subsets of the training data. Bootstrapping involves random sampling with replacement from the original data. The number of subsets created is usually determined in advance.
3) Base model training: For each created subset, train the same or similar classification model, such as a Decision Tree, on that subset.
4) Prediction with basic models: Use each basic model to make predictions on the data to be tested or test data.
5) Combining prediction results: Combine prediction results from all base models. In classification, majority voting is often used to determine the final classification label. In other words, the class most frequently selected by the basic models becomes the prediction label.
6) Final result: The result of majority voting is the final class prediction given by the Bagging model.
7) Model evaluation: Evaluate the performance of the Bagging model using evaluation metrics such as accuracy, precision, recall, or $F1$-score on test data.

F. Random Forests

The random forest is a classification algorithm that employs a collection of decision trees to generate predictions. It works by creating multiple decision trees and combining their predictions through voting [29], [30]. The Random Forest Classifier has several of important benefits, including the capacity to tolerate imbalanced data, the tolerance for over-fitting using various tree ensembles, the capacity to handle categorical and numerical data without the need for special transformations, and the capacity to gauge the weight of features in decision-making.

However, there are drawbacks that should be taken into consideration, such as the high computational resource consumption caused by using numerous decision trees, as well as the lack of inter-pretability, which makes prediction results challenging to intuitively explain, particularly in situations that call for a thorough understanding of the relationships between features. The Random Forests algorithm application process generally involves the following steps:

1) Select number of trees (n_estimators): The first step is to decide how many decision trees to use in the Random Forest ensemble.
2) Bootstrap sampling: For each tree in the ensemble, take a random sample (with replacement) from the training data. This produces a subset of the data that is used to train each tree.
3) Constructing a decision tree: Create a decision tree for each tree in the ensemble by doing the following steps: a.
   a) For each node split, pick a random subset of characteristics to consider.
   b) Commence the process by identifying the root node and partitioning the data according to the feature that offers the most effective means of distinction, which could be determined by evaluating either the Gini impurity or the information gain.
   c) Continue the process of separating until it arrives at a condition that brings it to a halt, such as attaining a predetermined maximum depth or obtaining a minimum number of samples at a terminal node.
   d) Predict with each tree: Once all decision trees have been trained, use each tree to make predictions on test data or unseen data.
   e) Results aggregation: The aggregation of outcomes from individual trees is contingent upon the specific task at hand. In the case of classification, the prevailing approach is typically to employ the majority voting method. The class that is chosen by many trees is subsequently deemed the ultimate prediction for the class.

III. Result

As an initial analysis, feature correlation was visualized using a heat map. The results of this visualization can be seen in Fig. 2. Based on Fig. 2, it is known that the correlation between features is very low. This means that changes to one feature will not majorly affect other features. Apart from that, the low correlation between these features also indicates that the redundant features are not redundant with each other.

The experiments in this research were carried out using three different classification methods. The aim of using these three different methods is to analyze which method is the best as a malaria prediction model. Furthermore, the methods used in this research are Decision Tree, Bagging, and Random Forests. The dataset used in this research is divided into 70 % training set and 30 % testing set. The division of this dataset is based on the proportion of training data and testing data, which are commonly used in classification tasks.

A. Experiment Results of Decision Tree Classifier

The first experiment used the Decision Tree method to build a prediction model. The parameter of the model built is the maximum depth of the tree. When setting

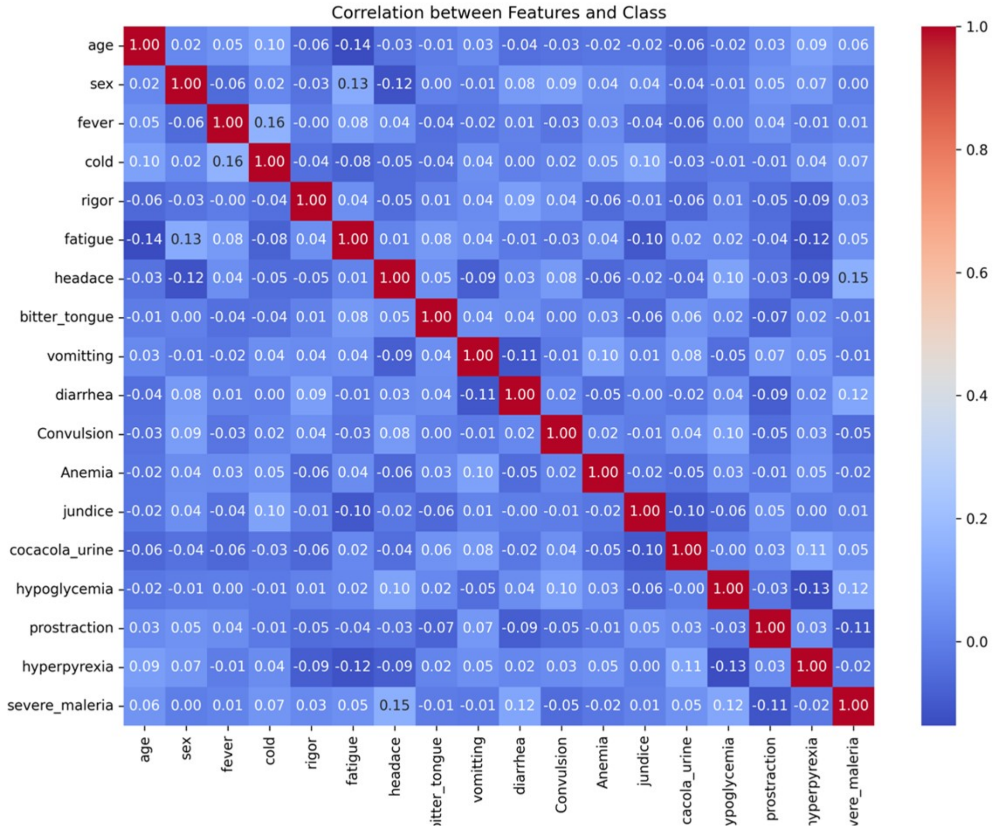

Fig. 2. Features correlation visualization.

the maximum tree depth, it will set how far the decision tree can grow. This is one of the hyper-parameters that can be tuned during the Decision Tree model training process.

Setting the maximum tree depth can help in avoiding over-fitting. The depth of the tree can be adjusted according to the characteristics of the data and research objectives. If the depth of the tree is too deep, then the model tends to over-fit; if it is too shallow, then the model may not be able to capture complex patterns in the data.

In this study, seven different tree depth values were used, and these values were determined based on tree depth values that are widely used in previous studies. Furthermore, the result of the experiment using the decision tree method can be seen in Table 2. Based on the results in Table 2, the best accuracy is obtained when the tree depth is 20, compared to shallower or deeper tree depths. This means that in the decision tree model, the choice of tree depth does not always follow the rule ”the deeper, the better” or ”the deeper, the more accurate.”

The optimal tree depth depends on the data used in the research. This means that in this research, the optimal tree depth value is 20. Furthermore, the average accuracy of the decision Tree model is 56.93 %.

Table 2. Experiment Results Using Decision Tree Classifier

Maximum Tree Depth	**Accuracy**
10	57.35%
15	55.88%
20	58.82%
25	58.82%
30	57.35%
40	55.88%
50	54.41%
Average	**56.93%**

B. *Experiment Results of Bagging Classifier*

The second experiment was carried out using the Bagging (Bootstrap Aggregating) method. In the Bagging method for classification, ”base model” refers to the basic model used in the ensemble. An ensemble is a combination of several models used to improve prediction or classification performance compared to using one single model.

As for the experiments carried out in this research, the base model used was Decision Tree. The parameter used is the number of base models, which are then called n-estimators. The n-estimators in the bootstrap aggregating (Bagging) classification method refer to the number of base models that will be used in the ensemble. In the context of the Bagging method, the ensemble consists of several base models which are generated by training the base model using different datasets randomly using the bootstrapping technique.

Furthermore, the experiments in this research used 7 different values of n-estimators, to find out the optimal value. The results of this experiment can be seen in Table 3. Based on the results in Table 3, it is known that the highest accuracy is obtained when the n-estimators value is 40. This means that an ensemble consisting of 40 decision trees produced using the Bagging technique can provide the best or most accurate performance, compared to the n-estimators value, lower or higher. These results also show that the ensemble with 40 base Decision Tree models is effective in reducing variance and increasing prediction accuracy.

Table 3. Experiment Results using Bagging Classifier

n-estimators	Accuracy
10	66.18%
20	64.71%
30	67.65%
40	69.12%
50	64.71%
75	66.18%
100	66.18%
Average	**66.39%**

C. Experiment Results of Random Forests Classifier

The last experiment was carried out using a Random Forests classifier. Furthermore, this experiment was using two hyper-parameters. Namely the k value in k-fold validation and the number of decision trees in the Random Forests classifier. The experiment employed k-values of 5, 10, 15, and 20, and decision tree values of 100, 200, 300, 400, and 500. The aim of using different k-fold and tree values is to determine the best number of trees for the classification process in this research. Furthermore, the experiment results of this research can be seen in Table 4, where *Avg.* is the abbreviation of average accuracy.

Table 4. Experiment Results using Random Forests Classifier

k-value	**Tree values**					**Avg.**
	100	**200**	**300**	**400**	**500**	
5	68%	67%	67%	68%	68%	67.6%
10	76%	76%	73%	73%	73%	74.2%
15	77%	73%	77%	77%	77%	76.2%
20	82%	82%	82%	82%	82%	82.0%
Avg.	75.8%	74.5%	74.8%	75.0%	75.0%	**75%**

Table 4 shows that the best results were obtained using a k-value of 20. This shows that the highest accuracy in k-fold cross-validation was obtained using 20 folds (k = 20). This indicates that the model tested in this experiment tends to provide the best results when tested with a cross-validation technique that divides the data into 20 subsets. This indicates the model's ability to generalize well and is stable in the face of different test data, which is critical for measuring the overall performance of the model.

Meanwhile, the number of folds in k-fold is 20, indicating that the Random Forest model reaches a saturation point in increasing accuracy when it reaches 100 trees. This means that adding further trees does not result in a significant improvement in model performance because a model with 100 trees is already powerful enough to cope with data variability. Furthermore, the average accuracy of the model using the Random Forests classifier is 75 %.

As a comparison, the best results in this study are compared with the results from previous studies in Table 5. Table 5 shows that the Random Forests method used for classifying Malaria data provides better results than the results in previous studies. The methods used in previous research were support vector machine (SVM) and CART.

Table 5. Comparison of the Result on This Research with Previous Researches.

Author	Method	Accuracy
[1]	Min Max-SVM	64%
[2]	CART	77%
[3]	Min Max-CART	56.2%
This research	**Robust Scaler-Random Forest**	**82%**

This may be due to Random Forests' ability to handle data complexity and variability in the Malaria dataset using ensemble learning, which reduces the possibility of over-fitting. Random Forests may also be better at adapting to class imbalance, thereby improving performance in the case of imbalanced data.

Additionally, Random Forests provide more powerful models by combining several decision trees, which can capture more subtle patterns in your data. The choice of classification method should always be adjusted to the specific characteristics of the dataset and research objectives, and these results show that in the Malaria data classification process, Random Forests are an effective choice to increase accuracy in the classification task.

D. Comparison of the Results of Three Classifiers

Fig. 3 shows a comparison of the highest accuracy and average accuracy of the three methods used in this research. It shows that the Bagging method can provide higher accuracy both in terms of highest accuracy, compared to the Decision Tree method.

This happens because in the Bagging method, several Decision Trees are created using subsamples of the training data, and then the prediction results from these trees are combined. This reduces the variance in the model, makes it more stable, and is more likely to avoid over-fitting.

In addition, Bagging can overcome high variation and complexity in the dataset by presenting several variations of the training data. Combining prediction results from multiple trees, Bagging produces more consistent and accurate predictions, providing better performance than a single Decision Tree model in the experiments conducted.

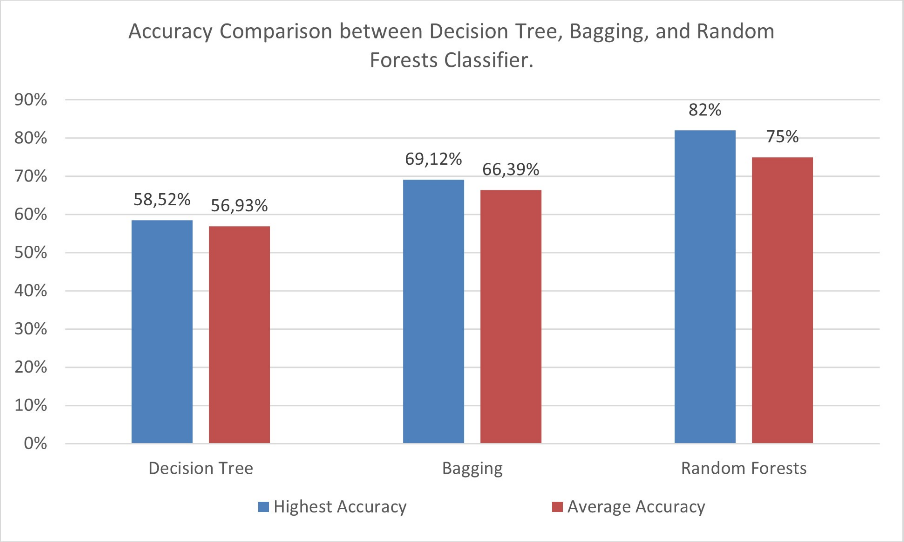

Fig. 3. Accuracy comparison between three method.

IV. Discussion

This research shows that the Random Forests method provides the highest accuracy and a higher average accuracy than the Decision Tree and Bagging methods. This happens because Random Forests is an ensemble method that combines the concepts of the Bagging method with the use of subsequences in selecting variables (features). This results in more diverse decision trees because each tree only looks at a small portion of the available features. This variability, along with the weighting used in combining prediction results, makes Random Forests more effective in reducing over-fitting and increasing its ability to generalize to never-before-seen data. In this way, Random Forests can achieve higher accuracy than Decision Trees, which may be more prone to over-fitting.

Additionally, Random Forests also have a built-in ability to assess the importance of each feature in the decision-making process. In other words, Random Forests can provide information about which features contribute most to making accurate predictions. By combining these features efficiently, Random Forests can achieve better results than Bagging or Decision Trees, which may have different mechanisms for assessing feature importance.

Therefore, the experimental results in this study show that Random Forests are a strong choice for improving prediction accuracy. Thus, the highest accuracy of the Random Forests method is obtained when the k value used is k = 20. This means increasing the folds (k-value) or subsets of data used to train and test the model. This optimizes data usage, reduces variability in results, and helps avoid over-fitting to the training data that may occur with fewer folds. With higher k, model evaluation becomes more stable and can provide more consistent estimates of overall model performance on the dataset.

Meanwhile, using a different number of decision trees provides slightly different or even the same accuracy. This happens because, in Random Forests, each tree is trained independently with bootstrapped data samples and random feature selection. When the prediction results from many such trees are combined through a voting (for classification) or averaging (for regression) process, the tendency of the different trees to offset each other's errors reduces variance and over-fitting.

Thus, even with different numbers of trees, the resulting models are still able to provide similar results because this ensemble process allows the models to handle variations in the data well. However, more trees usually provide more stable estimates and can increase robustness to over-fitting.

V. Conclusion

In this study, a system to predict malaria was developed using a dataset from Nigeria that had 18 characteristics. Three approaches are used in this prediction system, to conclude what method is most suitable for this research. The results compared in this research are the highest accuracy and average accuracy produced by each method.

Furthermore, the results of the experiments carried out show that the Bagging method can provide higher accuracy than the decision trees method. This happens because in the Bagging method, several Decision Trees are created using subsamples of the training data, and then the prediction results from these trees are combined. This reduces the variance in the model, makes it more stable, and is more likely to avoid over-fitting.

Because each tree only looks at a small portion of the available features, this results in a more diverse decision tree. Random Forests have more variations and weights used to combine prediction results, which makes them better at reducing over-fitting and generalizing to never-before-seen data. Thus, random forests can achieve higher levels of accuracy than single Decision Trees, which may be more prone to over-fitting.

By systematically varying hyper-parameters, this research concludes an optimal accuracy of 82 % was achieved at a k-value of 20, demonstrating the effectiveness and robustness of the Random Forests method.

Therefore, random forests are a good choice to improve prediction accuracy, according to the experimental results of this research. Overall, this research provides a promising approach to predicting malaria occurrence based on symptomatic presentation, which can aid in preventive strategies and contribute to global efforts in combating this significant health concern.

REFERENCES

[1] N. G. Ramadhan and A. Khoirunnisa, “Klasifikasi data malaria menggunakan metode support vector machine,” *J. MEDIA Inform. BUDIDARMA*, vol. 5, no. 4, p. 1580, 2021, doi: 10.30865/mib.v5i4.3347.

[2] K. K. Republik Indonesia, “Wilayah-wilayah endemis malaria tinggi di Indonesia,” [Online]. Available: https://p2pm.kemkes.go.id/publikasi/artikel/wilayah-wilayah-endemis-malaria-tinggi-di-indonesia

[3] K. Y. Tai and J. Dhaliwal, “Machine learning model for malaria risk prediction based on mutation location of large-scale genetic variation data,” *J. Big Data*, vol. 9, no. 1, p. 85, 2022, doi: 10.1186/s40537-022-00635-x.

[4] Y. A. A. Et. Al., “Malaria prediction model using machine learning algorithms,” *Turk. J. Comput. Math. Educ. TURCOMAT*, vol. 12, no. 10, pp. 7488−7496, 2021, doi: 10.17762/turcomat.v12i10.5655.

[5] Y. A. Adamu and J. Singh, “Hybrid machine learning algorithm for prediction of malaria,” in *Proceedings of Fourth International Conference on Computing, Communications, and Cyber-Security*, vol. 664, S. Tanwar, S. T. Wierzchon, P. K. Singh, M. Ganzha, and G. Epiphaniou, Eds., in Lecture Notes in Networks and Systems, vol. 664. , Singapore: Springer Nature Singapore, 2023, pp. 413−423. doi: 10.1007/978-981-99-1479-1_31.

[6] Y. W. Lee, J. W. Choi, and E.-H. Shin, “Machine learning model for predicting malaria using clinical information,” *Comput. Biol. Med.*, vol. 129, p. 104151, 2021, doi: 10.1016/j.compbiomed.2020.104151.

[7] M. Mariki, E. Mkoba, and N. Mduma, “Combining clinical symptoms and patient features for malaria diagnosis: Machine learning approach,” *Appl. Artif. Intell.*, vol. 36, no. 1, p. 2031826, 2022, doi: 10.1080/08839514.2022.2031826.

[8] S. S. Yadav, V. J. Kadam, S. M. Jadhav, S. Jagtap, and P. R. Pathak, “Machine learning based malaria prediction using clinical findings,” in *2021 International Conference on Emerging Smart Computing and Informatics (ESCI)*, Pune, India: IEEE, Mar. 2021, pp. 216−222. doi: 10.1109/ESCI50559.2021.9396850.

[9] G. Shekar, S. Revathy, and E. K. Goud, “Malaria detection using deep learning,” in *2020 4th International Conference on Trends in Electronics and Informatics (ICOEI)*, Tirunelveli, India: IEEE, Jun. 2020, pp. 746−750. doi: 10.1109/ICOEI48184.2020.9143023.

[10] J. B. Awotunde, R. G. Jimoh, I. D. Oladipo, and M. Abdulraheem, “Prediction of malaria fever using long-short-term Memory and big data,” in *Information and Communication Technology and Applications*, vol. 1350, S. Misra and B. Muhammad-Bello, Eds., in Communications in Computer and Information Science, vol. 1350. , Cham: Springer International Publishing, 2021, pp. 41−53. doi: 10.1007/978-3-030-69143-1_4.

[11] B. Muhammad and A. Varol, “A Symptom-based machine learning model for malaria diagnosis in Nigeria,” in *2021 9th International Symposium on Digital Forensics and Security (ISDFS)*, Elazig, Turkey: IEEE, Jun. 2021, pp. 1−6. doi: 10.1109/ISDFS52919.2021.9486315.

[12] Y. P. Bria, C.-H. Yeh, and S. Bedingfield, “Machine learning classifiers for Symptom-based malaria prediction,” in *2022 International Joint Conference on Neural Networks (IJCNN)*, Padua, Italy: IEEE, Jul. 2022, pp. 1−6. doi: 10.1109/IJCNN55064.2022.9891945.

[13] R. Irmanita, S. S. Prasetiyowati, and Y. Sibaroni, “Classification of malaria complication using CART (classification and regression tree) and Naive Bayes,” *J. RESTI Rekayasa Sist. Dan Teknol. Inf.*, vol. 5, no. 1, pp. 10−16, 2021, doi: 10.29207/resti.v5i1.2770.

[14] S. Kuzhaloli, S. Thenappan, P. T, V. Nivedita, M. Mageshbabu, and S. Navaneethan, “Identification of malaria disease using machine learning models,” in *2023 Fifth International Conference on Electrical, Computer and Communication Technologies (ICECCT)*, Erode, India: IEEE, Feb. 2023, pp. 1−4. doi: 10.1109/ICECCT56650.2023.10179665.

[15] J. B. Awotunde, S. Misra, F. E. Ayo, A. Agrawal, and R. Ahuja, “Hybridized support vector machine and Adaboost technique for malaria diagnosis,” in *Frontiers of ICT in Healthcare, vol. 519, J. K. Mandal and D. De, Eds., in Lecture Notes in Networks and Systems*, vol. 519. , Singapore: Springer Nature Singapore, 2023, pp. 25−38. doi: 10.1007/978-981-19-5191-6_3.

[16] M. O. Arowolo, M. O. Adebiyi, C. T. Nnodim, S. O. Abdulsalam, and A. A. Adebiyi, “An adaptive genetic algorithm with recursive feature elimination approach for predicting malaria vector gene expression data classification using support vector machine Kernels,” *Walailak J. Sci. Technol. WJST*, vol. 18, no. 17, 2021, doi: 10.48048/wjst.2021.9849.

[17] N. O. Adeboye, O. V. Abimbola, and S. O. Folorunso, “Malaria patients in Nigeria: Data exploration approach,” *Data Brief*, vol. 28, p. 104997, 2020, doi: 10.1016/j.dib.2019.104997.

[18] H. Gunduz, “Deep learning-based Parkinson’s disease classification using vocal feature sets,” *IEEE Access*, vol. 7, pp. 115540−115551, 2019, doi: 10.1109/ACCESS.2019.2936564.

[19] N. G. Ramadhan and I. Atastina, “Neural network on stock prediction using the stock prices feature and Indonesian financial news titles,” *Int. J. Inf. Commun. Technol. IJoICT*, vol. 7, no. 1, pp. 54−63, 2021, doi: 10.21108/ijoict.v7i1.544.

[20] A. Khoirunnisa, Adiwijaya, and A. A. Rohmawati, “Implementing principal component analysis and multinomial logit for cancer detection based on microarray data classification,” in *2019 7th International Conference on Information and Communication Technology (ICoICT)*, Kuala Lumpur, Malaysia: IEEE, Jul. 2019, pp. 1−6. doi: 10.1109/ICoICT.2019.8835320.

[21] D. Singh and B. Singh, “Investigating the impact of data normalization on classification performance,” *Appl. Soft Comput.*, vol. 97, p. 105524, 2020, doi: 10.1016/j.asoc.2019.105524.

[22] D. Singh and B. Singh, “Feature wise normalization: An effective way of normalizing data,” *Pattern Recognit.*, vol. 122, p. 108307, 2022, doi: 10.1016/j.patcog.2021.108307.

[23] K. V. A. Reddy, S. R. Ambati, Y. S. Rithik Reddy, and A. N. Reddy, “AdaBoost for Parkinson’s disease detection using robust scaler and SFS from acoustic features,” in *2021 Smart Technologies, Communication and Robotics (STCR)*, Sathyamangalam, India: IEEE, Oct. 2021, pp. 1−6. doi: 10.1109/STCR51658.2021.9588906.

[24] B. Charbuty and A. Abdulazeez, “Classification based on decision tree algorithm for machine learning,” *J. Appl. Sci. Technol. Trends*, vol. 2, no. 01, pp. 20−28, 2021, doi: 10.38094/jastt20165.

[25] H. Zhou, J. Zhang, Y. Zhou, X. Guo, and Y. Ma, "A feature selection algorithm of decision tree based on feature weight," *Expert Syst. Appl.*, vol. 164, p. 113842, 2021, doi: 10.1016/j.eswa.2020.113842.

[26] Md. S. Bin Alam, M. J. A. Patwary, and M. Hassan, "Birth mode prediction using Bagging ensemble classifier: A Case Study of Bangladesh," in *2021 International Conference on Information and Communication Technology for Sustainable Development (ICICT4SD)*, Dhaka, Bangladesh: IEEE, Feb. 2021, pp. 95−99. doi: 10.1109/ICICT4SD50815.2021.9396909.

[27] H. Jafarzadeh, M. Mahdianpari, E. Gill, F. Mohammadimanesh, and S. Homayouni, "Bagging and boosting ensemble classifiers for classification of multispectral, hyperspectral and PolSAR data: A comparative evaluation," *Remote Sens.*, vol. 13, no. 21, p. 4405, 2021, doi: 10.3390/rs13214405.

[28] S. E. Roshan and S. Asadi, "Improvement of Bagging performance for classification of imbalanced datasets using evolutionary multi-objective optimization," *Eng. Appl. Artif. Intell.*, vol. 87, p. 103319, 2020, doi: 10.1016/j.engappai.2019.103319.

[29] Y. Liu, Y. Wang, and J. Zhang, "New machine learning algorithm: Random Forest," in *Information Computing and Applications*, vol. 7473, B. Liu, M. Ma, and J. Chang, Eds., in Lecture Notes in Computer Science, vol. 7473. , Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 246−252. doi: 10.1007/978-3-642-34062-8_32.

[30] N. G. Ramadhan, A. Adiwijiya, W. Maharani, and A. A. Gozali, "Prediction of diabetes mellitus in the upcoming year using SMOTE and random forest," in *2023 International Conference on Data Science and Its applications (ICoDSA)*, Bandung: IEEE, Aug. 2023.