JIKO (Jurnal Informatika dan Komputer)
Vol. 7, No. 2, August 2024, pp. 88-92
DOI:10.33387/jiko.v7i2.8496

Accredited KEMENDIKBUD RISTEK, No.105/E/KPT/2022
p-ISSN: 2614-8897
e-ISSN: 2656-1948

CLASSIFICATION OF DENGUE FEVER DISEASE USING A MACHINE LEARNING-BASED RANDOM FOREST ALGORITHM

Arif Fitra Setyawan[1], Amelia Devi Putri Ariyanto[2] , Fari Katul Fikriah [3]

[1,2,3]Program Studi Sistem dan Teknologi Informasi, Fakultas Keperawatan Bisnis dan Teknologi, Universitas Widya Husada, Indonesia
[*]Email: [1]ariffitra.setyawan@gmail.com, [2]ameliadev26@gmail.com, [3]farichatulfikriyah45@gmail.com

(Received: 22 July 2024, Revised: 29 July 2024, Accepted: 6 August 2024)

Abstract

Dengue Hemorrhagic Fever (DHF) is a tropical disease that often results in high morbidity and mortality rates. Early diagnosis of DHF is crucial to mitigate its adverse effects. However, manual diagnostic processes are often inefficient and prone to errors. This study aims to develop a DHF classification model using the Random Forest algorithm, which is expected to assist in the early diagnosis of this disease. The methodology used in this research is CRISP-DM (Cross-Industry Standard Process for Data Mining), which includes the stages of Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment. Data was obtained from kaggle.com, and during the Data Preparation stage, missing values were removed, categorical features were encoded, data was normalized, and split into training and testing sets. The research results show that the Random Forest model has an accuracy of 88.5%, precision of 88.2%, recall of 65.2%, F1-score of 74.9%, and ROC AUC of 0.810. Feature importance analysis revealed that the Gender_Male and Body_Pain features have the largest contributions in DHF classification. Although the model demonstrated high accuracy and precision, the lower recall value indicates that some positive cases were missed, requiring further improvements. The Random Forest can be used as a tool for early DHF diagnosis, but further adjustments are necessary to enhance its performance. This research provides insights into the contributing factors for DHF diagnosis and the practical application potential of this model in medical decision support systems.

Keywords: *CRISP-DM, Dengue Hemorrhagic Fever, Classification, Machine Learning, Random Forest*

**Corresponding Author: Arif Fitra Setyawan*

1. INTRODUCTION

Diseases related to the environment remain a public health issue to this day. One of the diseases caused by poor environmental sanitation conditions is dengue fever (Dengue Hemorrhagic Fever or DHF) [1]. Dengue fever is an infectious disease caused by the dengue virus. Dengue Hemorrhagic Fever (DHF) is an epidemic that affects various countries worldwide, with over 500,000 cases reported annually. [2]. This has become a significant global health issue, especially in tropical and subtropical regions. The disease spreads rapidly and is often fatal because many patients die due to delayed treatment. [3]. According to the World Health Organization (WHO), Dengue Hemorrhagic Fever (DHF) or Dengue Fever (DF) is a disease caused by the bite of an Aedes mosquito infected with one of the four types of dengue virus, presenting with clinical manifestations such as fever, muscle and/or joint pain, along with leukopenia, rash, lymphadenopathy, thrombocytopenia, and hemorrhagic diathesis [4]. Early detection and management of this disease are crucial to reducing its impact. The diagnosis of dengue fever often relies on laboratory test results and clinical symptoms. This can be a challenging task that requires deep medical knowledge. Additionally, the early symptoms of dengue fever often resemble those of other illnesses, such as the flu or other infectious diseases, making it difficult to make an accurate diagnosis.

Information technology and machine learning have unlocked significant potential for supporting disease diagnosis. Data mining is the process of using statistical, mathematical, and artificial intelligence techniques to extract and identify information and patterns from large datasets. Its primary goal is

information extraction, particularly in the form of classification. [5]. Although various terms such as knowledge mining or knowledge discovery are used to refer to data mining, this concept has become a popular method in society. Data mining is also known as knowledge extraction, pattern analysis, data archaeology, information harvesting, and pattern discovery. By uncovering the facts stored within data, data mining generates valuable knowledge through deep analysis processes.

Classification in data mining is a technique that groups data into predefined categories. The goal is to predict the class value of unknown objects by identifying patterns that represent data classes. This process involves validating the model using training data for model creation and testing the model's accuracy with test data. Classification enables the prediction of the name or value of data objects by utilizing identified patterns [6].

To detect early symptoms of diabetes promptly, it is necessary to develop a model and classification using data mining. This research will utilize the Cross-Industry Standard Process for Data Mining (CRISP-DM) methodology and the Random Forest algorithm [7]. The CRISP-DM (Cross Industry Process Model for Data Mining) methodology is a well-established approach that continues to be widely recognized in data mining projects through the application of machine learning algorithms [8] [9]. Random Forest is a machine learning algorithm used to develop decision trees. It can be considered as a combination of multiple decision trees [10]. Predictions from the Random Forest algorithm are obtained through the majority vote of each individual decision tree (voting process for classification and averaging for regression) [11].

In previous research on Random Forest, it was proven effective in classifying and detecting stroke symptoms, with the initial data divided into 80% for training and 20% for testing. Validation was performed using cross-validation, achieving a training score of 96%, an accuracy of 95%, and an AUC value of 0.80, indicating strong model performance [12]. Research using the Random Forest algorithm has also been conducted to predict diabetes, with results indicating that the algorithm can predict diabetes with high performance. This algorithm has proven to be highly reliable as a reference for developing predictive models for similar cases, as reflected by an AUC value reaching 100%. [13]. Research comparing the Random Forest algorithm and SVM showed that Random Forest outperformed with an accuracy of 88.2% on the test data and 98.8% on the training data, which is better than SVM. After hypertuning the SVM algorithm, its accuracy improved to 81%, approaching the accuracy level of Random Forest [14]. The Random Forest machine learning algorithm has proven effective in classifying data and predicting outcomes based on patterns identified within the data.

The main objective of this research is to develop a classification model with a high accuracy rate in recognizing dengue fever cases. This model is expected to accurately differentiate between dengue and non-dengue cases based on patient data. Through this study, a reliable classification model is hoped to be produced that can be implemented in healthcare facilities, supporting better clinical decision-making and improving overall public health. This research also aims to identify the most significant clinical features in dengue classification, which can assist in medical decision-making. Thus, it is anticipated that this study will contribute significantly to enhancing the quality of dengue diagnoses and assist healthcare professionals in providing faster and more accurate treatment.

2. RESEARCH METHOD

This research applies the CRISP-DM (Cross-Industry Standard Process for Data Mining) methodology, which includes six main phases: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment [15] [16].

2.1. Business Understanding

This initial step focuses on interpreting business objectives and requirements, which are then translated into insights to define the main problem that can be addressed through data mining [17]. In the Data Understanding phase, clinical data related to Dengue Fever (DF) is sourced from the dataset available on Kaggle.com. This dataset includes essential information such as symptoms, vital signs, laboratory results, and final diagnosis. The initial data exploration process involves understanding the data distribution, identifying missing values, and detecting outliers that may affect the analysis results. Descriptive analysis is performed to gain initial insights into the data characteristics, such as the number of cases, the demographic distribution of patients, and common symptom patterns. A deep understanding of the data structure and quality is crucial to ensure that the data is ready for the preprocessing phase and the development of machine learning models.

2.2. Data Understanding

In this phase, data visualization is performed to understand the data and clean it by addressing missing data or removing problematic features, aiming to produce a better and more generalizable machine learning model [18]. Clinical data related to Dengue Fever (DF) is sourced from a dataset available on Kaggle.com. This dataset includes crucial information such as symptoms, vital signs, laboratory results, and final diagnoses. The initial data exploration process involves understanding data distribution, identifying missing values, and detecting outliers that may impact the analysis results. Descriptive analysis is performed to gain preliminary insights into data characteristics, including the number of cases, patient demographic distribution, and common symptom patterns. A

thorough understanding of the data's structure and quality is essential to ensure the data is ready for the preprocessing stage and the development of machine learning models.

2.3. Data Preparation

Data preparation is a crucial stage in the CRISP-DM methodology aimed at ensuring the data for machine learning models is clean, relevant, and in the appropriate format. In this study, the steps taken include removing missing values to prevent disruption to the model, and encoding categorical features like gender using one-hot encoding to ensure compatibility with the Random Forest algorithm. Irrelevant features are also removed to reduce complexity and improve prediction accuracy. Additionally, data normalization is performed if necessary, to ensure numerical features fall within a uniform range, aiding faster convergence of the algorithm. The dataset is then split into training (80%) and testing (20%) data to prevent overfitting and ensure good model generalization. Finally, balancing techniques such as oversampling or undersampling are applied if class imbalance is present, to prevent bias towards the majority class. These steps ensure optimal data for training the Random Forest model, contributing to better performance in classifying Dengue Fever.

2.4. Modeling

In the Modeling phase, the Random Forest algorithm is used to develop a classification model for Dengue Fever (DF). Initially, the dataset is divided into training and testing sets using k-fold cross-validation to prevent overfitting. The model is trained on the training data, with hyperparameter tuning performed via grid search or random search to optimize its performance. After training, the model is evaluated using the testing data with metrics such as accuracy, precision, recall, F1-score, and AUC-ROC to assess its effectiveness. Feature importance analysis is also conducted to identify the most significant features for DF classification.

2.5. Evaluation

In the Evaluation phase, the classification model for Dengue Fever (DF) is assessed to measure its performance in predicting DF cases using relevant metrics. After training and optimizing the Random Forest model through hyperparameter tuning, we test the model on a separate testing dataset. The evaluation involves using several metrics, including accuracy, precision, recall, F1-score, and Area Under the Receiver Operating Characteristic Curve (AUC-ROC).

3. RESULT AND DISCUSSION

This study employs the Random Forest algorithm to classify Dengue Fever (DF) using the CRISP-DM (Cross-Industry Standard Process for Data Mining) methodology. The process begins with the Business Understanding phase, which establishes the goal of developing a predictive model to assist in the early diagnosis of DF. In the Data Understanding phase, data from kaggle.com is analyzed to understand distribution, identify missing values, and explore feature relationships.

The dataset from Kaggle consists of 5000 entries and includes seven variables: Temperature, Platelet_Count, White_Blood_Cell_Count, Body_Pain, Rash, Gender, and Infected. It provides detailed health parameters related to Dengue Fever. The average temperature is 98.49°F, with a range from 97.50°F to 99.50°F. Platelet count averages 149,631.33 cells, varying between 73,266.89 and 220,581.10 cells, while the white blood cell count has a mean of 7002.35 cells, with a range from 3077.60 to 10428.91 cells. Body pain is present in all entries, whereas rash is found in 19.6% of the cases. Gender distribution is nearly equal, with 2489 females and 2511 males. Additionally, 48.1% of the patients are infected with Dengue Fever.

In the Data Preparation phase, the data is cleaned by removing missing values, encoding categorical features, and splitting it into training and testing datasets. Data Preparation involves clearing missing values, encoding categorical features, and dividing the data into 80% training and 20% testing sets.

During the Modeling phase, the Random Forest algorithm is utilized to build the classification model, with hyperparameter tuning conducted via Grid Search to find the optimal parameter combination. The model is trained on the training data and tested on the testing data to evaluate its performance.

Model evaluation was conducted using metrics such as accuracy, precision, recall, F1-score, ROC AUC, and K-Fold Cross-Validation. The model evaluation results is shown in the following table;

Table 1. Model Evaluation Results Table

Metric	Value
Accuracy	88.5%
Precision	88.2%
Recall	65.2%
F1-Score	0.749
ROC AUC	0.810
Mean Accuracy	88.5%
Standard Deviation	0

The Random Forest model achieves an accuracy of 88.5%, indicating that it correctly classifies about 88.5% of the tested data. This high accuracy suggests the model's potential for DF classification. A precision of 88.2% means that 88.2% of the model's positive predictions are truly positive. The recall of 65.2% shows that the model detects 65.2% of all actual positive cases. Although this recall value is lower than precision, it is acceptable depending on the clinical context and application priorities. Improving recall is crucial to minimize missed positive cases. The F1-score of 0.749 balances precision and recall, reflecting

a reasonable performance in balancing true positive detection and minimizing false negatives. The ROC AUC value of 0.810 indicates that the model performs well in distinguishing between positive and negative classes, with a value above 0.8 indicating strong performance. The results of the k-fold cross-validation evaluation show that the model has an average accuracy of 88.5%, reflecting the model's ability to consistently make correct predictions across different folds of data. With a standard deviation of 0.0000, this indicates that the model's performance is highly stable and does not exhibit significant variability across the cross-validation folds. This consistency suggests that the model is not overly dependent on any specific subset of data, making it reliable for making predictions on new, similar data.

The results from the confusion matrix are shown in the following table 2:

Table 2. confusion matrix

		Actual values	
		Positive (1)	**Negative (0)**
Predicted values	**Positive (1)**	713	23
	Negative (0)	92	172

The model produces 713 true negatives (TN) and 172 true positives (TP). However, there are 23 false positives (FP) and 92 false negatives (FN). The higher number of FN compared to FP suggests that the model tends to overlook positive cases more frequently, indicating a potential area for further improvement.

The Feature Importance results are displayed in the following graph; here is the Feature Importance chart generated:

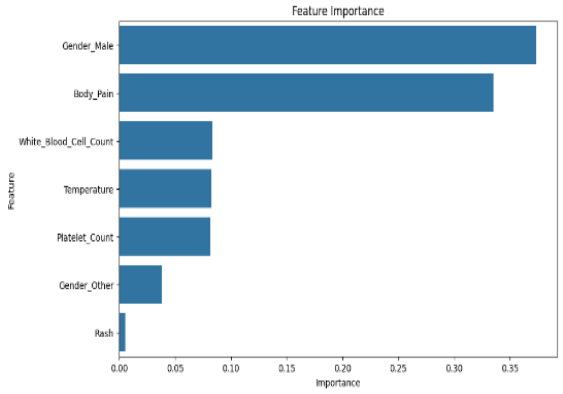

Figure 1. Feature Importance chart

Feature importance analysis reveals that Gender_Male and Body_Pain are the most significant contributors to the model, with importances of 37.4% and 33.5%, respectively. They are followed by White_Blood_Cell_Count (8.3%) and Temperature (8.2%). The feature Rash has the lowest contribution at 0.6%. This analysis helps identify which features most impact the model's predictions and can provide further insights for medical research.

4. CONCLUSION

This study developed a classification model for Dengue Fever (DF) using the Random Forest algorithm and the CRISP-DM methodology. The evaluation results indicate that the model performs well with an accuracy of 88.5%, precision of 88.2%, recall of 65.2%, F1-score of 74.9%, and ROC AUC of 0.810. The Data Preparation phase involved removing missing values, encoding categorical features, eliminating irrelevant features, normalizing data, and splitting the data into training and testing sets. Feature importance analysis revealed that Gender_Male and Body_Pain are the most significant contributors. Despite the high accuracy and precision, the lower recall indicates that some positive cases were missed, suggesting the need for further improvement. This study provides insights into the factors contributing to DF diagnosis and demonstrates the potential practical application of the model in medical decision support systems.

5. REFERENCES

[1] I. A. Dania, "GAMBARAN PENYAKIT DAN VEKTOR DEMAM BERDARAH DENGUE (DBD)," *Jurnal Warta*, 2016.

[2] T. Firman dan A. Ahmedika, "Diagnosa Penyakit Demam Berdarah Dengue (DBD) menggunakan Metode Learning Vector Quantization (LVQ)," *JISKa*, p. 193 – 201, 2020.

[3] D. A. Reza, N. N. Yuki dan S. Wahyuningsih, "KLASIFIKASI PROBABILISTIC NEURAL NETWORK (PNN) PADA DATA DIAGNOSA PENYAKIT DEMAM BERDARAH DENGUE (DBD) TAHUN 2018," dalam *Prosiding Seminar Nasional Matematika, Statistika, dan Aplikasinya 2019*, Samarinda, 2019.

[4] A. H. Chainur, A. M. Moch dan A. Prahutama, "KLASIFIKASI DIAGNOSA PENYAKIT DEMAM BERDARAH DENGUE (DBD) MENGGUNAKAN SUPPORT VECTOR MACHINE (SVM) BERBASIS GUI MATLAB," *JURNAL GAUSSIAN*, pp. 171-180 , 2017.

[5] Y. P. Arifin, N. F. Hari, M.Fakhrizal dan A. Nur, "Penerapan Data Mining Dalam Analisis Prediksi Kanker Paru Menggunakan Algoritma Random Forest," *Jurnal Ilmiah Teknik Informatika dan Komunikasi*, 2023.

[6] M. Idris, "IMPLEMENTASI DATA MINING DENGAN ALGORITMA NAÏVE BAYES UNTUK MEMPREDIKSI ANGKA KELAHIRAN," *Jurnal Pelita Informatika*, pp. 421-428, 2019.

[7] C. M. Ajeng, Rusdah, L. H. Law dan A. Dian, "DETEKSI DINI GEJALA AWAL PENYAKIT DIABETES MENGGUNAKAN ALGORITMA RANDOM FOREST," *Idealis: Indonesia Journal Information System* , pp. 165-171, 2023 .

[8] W. Y. Ayele, "Adapting CRISP-DM for Idea Mining: A Data Mining Process for Generating Ideas Using a Textual Dataset," *International Journal of Advanced Computer Science and Applications(IJACSA),* vol. 11, pp. 20-32, 2020.

[9] A. S. Jairo, J. L. C. Diana, F. U. I. Samir dan J. R. Coronado-Hernández, "Predictive models assessment based on CRISP-DM methodology for students performance in Colombia - Saber 11 Test,," *Procedia Computer Science,* vol. 198, pp. 512-517, 2022.

[10] P. Rifkie, Algoritma machine learning, Bandung: Informatika Bandung, 2021.

[11] A. Y. S. Taghfirul dan J. P. Wawan, "IMPLEMENTASI SELEKSI FITUR INFORMATION GAIN RATIO PADA ALGORITMA RANDOM FOREST UNTUK MODEL DATA KLASIFIKASI PEMBAYARAN KULIAH," *Dinamika Informatika,* pp. 41-49 , 2023.

[12] P. S. Ary, P. P. Dwi, P. P. Jojor dan R. B. Khairul, "Implementasi Algoritma Random Forest Dalam Klasifikasi Diagnosis Penyakit Stroke," *Jurnal Penelitian Rumpun Ilmu Teknik (JUPRIT),* pp. 155-164, 2023.

[13] Sriyanto dan R. S. Agiska, "Prediksi Penyakit Diabetes Menggunakan Algoritma Random Forest," *JURNAL TEKNIKA,* pp. 163-172 , 2023.

[14] B. Mahmin, B. H. Dinda dan S. Oloan, "KLASIFIKASI PENYAKIT STUNTING DENGAN MENGGUNAKAN ALGORITMA SUPPORT VECTOR MACHINE DAN RANDOM FOREST," *Jurnal TEKINKOM,* pp. 540-549, 2023 .

[15] Firmansyah dan Y. Agus, "Prediksi Penyakit Jantung Menggunakan Algoritma Random Forest," *Jurnal Minfo Polgan,* pp. 2239-2246 , 2023.

[16] S. Regina, P. Madalena, A. Mariana, R. Mariana dan P. Hugo, "Harnessing Data Mining to Predict Survival Outcomes in Patients with Hepatic Cirrhosis,," *Procedia Computer Science,* vol. 238, pp. 938-943, 2024.

[17] A. A. Zharfan, R. P. Satriawan, A. S. Darmawan, Ricko, W. Adi, S. Aris dan F. Wijayanto, "Prediction of Hotel Booking Cancellation using CRISP-DM," dalam *4th International Conference on Informatics and Computational Sciences (ICICoS)*, Semarang, 2020.

[18] P. Ana, F. Diana, N. Cristiana, A. António dan M. José, "Data Mining to Predict Early Stage Chronic Kidney Disease," *Procedia Computer Science,* vol. 177, pp. 562-567, 2020.