Journal of Artificial Intelligence and Engineering Applications

Website: https://ioinformatic.org/

15th February 2026. Vol. 5. No. 2; e-ISSN: 2808-4519

Improving the Accuracy of Early Diagnosis of Dengue Hemorrhagic Fever Based on Clinical Symptoms Using Random Forest

Suci Lestari [1*], Rudi Kurniawan [2], Bani Nurhakim [3], Ahmad Rifa'I [4], Ryan Hamonangan [5]

[1,2,3,4,5] *STMIK IKMI Cirebon*

sucilestari6391@gmail.com[1*], rudi226ikmi@gmail.com[2], baninurhakim@gmail.com[3], a.rifaaii1408@gmail.com[4], mr.ryansilalahi@gmail.com[5]

Abstract

The development of machine learning in the field of health provides important opportunities for improving the accuracy of disease diagnosis, including dengue hemorrhagic fever (DHF), which remains a major health problem in Indonesia. This study aims to develop an early DHF diagnosis model based on the Random Forest algorithm using clinical symptom data from patients at the Kosasih Group Clinic. The research was conducted using a quantitative approach through the CRISP-DM stages, which included data acquisition, validation, cleaning, and preprocessing, covering missing value handling, normalization, and class imbalance management using SMOTE. The dataset was then divided using stratified sampling to maintain class proportions, followed by training the Random Forest model optimized using Bayesian Optimization to obtain the best combination of hyperparameters. Performance evaluation was carried out using accuracy, precision, recall, F1-score, and ROC-AUC metrics, and validated using stratified k-fold cross-validation to ensure model stability. Model interpretability was analyzed using SHAP and LIME to identify the contribution of each clinical symptom to the prediction. The results showed that the model was able to provide high classification performance, with increased sensitivity to DHF cases after applying SMOTE and consistent interpretation of clinical symptoms such as fever, joint pain, and nausea. These findings confirm the potential of Random Forest as a reliable model to support the development of AI-based clinical decision support systems (CDSS) for early diagnosis of DHF in primary health care facilities.

Keywords: *Dengue Hemorrhagic Fever, Clinical Diagnosis, Machine Learning, Random Forest, SHAP*

1. Introduction

The development of information technology over the past decade has triggered significant transformations in various sectors, including healthcare services, which now increasingly rely on the use of data. Advances in artificial intelligence and machine learning enable data processing to be carried out with greater speed and accuracy, while also being able to manage large volumes of information, thereby strengthening data-driven decision-making practices [1]. In the medical field, the integration of information systems with clinical data analysis opens up vast opportunities to improve diagnostic accuracy, predict disease progression, and manage patients in a more integrated manner. This technology allows data on symptoms, laboratory test results, and electronic medical records to be analyzed through a systematic approach to produce clinical findings that are more relevant to healthcare professionals [2]. In line with the development of the concept of data-driven healthcare, algorithms such as Random Forest are increasingly being used to process complex and diverse clinical data, thereby providing stronger analytical support in modern medical practice.

Despite rapid advances in health technology, efforts to control infectious diseases such as dengue fever (DF) continue to face major obstacles at the global level. Various findings indicate a significant increase in DHF cases over the past decade, accompanied by an increasingly widespread but uneven pattern of distribution across regions [3][4]. This condition is in line with the results of surveillance studies and bibliometric analyses that note an increase in research intensity and outbreak frequency, especially in Southeast Asia and the Americas [5][6]. In Indonesia itself, the number of DHF cases still shows annual fluctuations influenced by environmental factors, population density, and the effectiveness of early detection in health services. These challenges become even more complex in primary health facilities such as the Kosasih Clinic, where DHF symptoms, which often resemble other fever-related illnesses, frequently cause delays or inaccuracies in diagnosis. In addition, limited medical resources and suboptimal utilization of clinical data for automated analysis further increase the potential for errors in clinical decision-making. Thus, the integration of Machine Learning is becoming increasingly relevant as an innovative approach to improve the accuracy of diagnosis based on available clinical symptom information.

Previous studies have shown significant progress in the application of machine learning algorithms to support the early diagnosis of dengue hemorrhagic fever (DHF). Findings by Leung et al. and Long et al. indicate that predictive models that utilize data from various sources

are able to improve the accuracy of disease incidence estimates more consistently [7][8]. On the other hand, a study conducted by da Silva Neto et al. confirms that Clinical Decision Support Systems (CDSS) that integrate clinical symptoms and laboratory test results with Machine Learning algorithms can improve diagnostic sensitivity and minimize the risk of clinical errors in medical practice [9]. Random Forest-based models and other ensemble approaches have also been proven capable of identifying key symptom features, including body temperature, joint pain, and platelet levels, as relevant early markers for the early detection of DHF [10][11]. Furthermore, the development of CDSS with a hybrid pipeline equipped with a high level of interpretability through the SHAP method strengthens the potential for effective model implementation in mid-level health facilities, especially those that need analytical support that is understandable and reliable for medical personnel [12].

Based on these conditions, this study focused on developing a dengue fever diagnosis model based on patient clinical symptoms at the Kosasih Clinic using the Random Forest algorithm. This study addresses the challenges of inaccurate initial diagnosis, limited use of local clinical data, and the need for a model that is not only accurate but also transparent and explainable.

1.1 Research Objectives

1. Develop a Random Forest-based DBD classification model using clinical symptom data;
2. Optimize model performance through SMOTE and Optuna;
3. Provide model interpretation using SHAP.

1.2 Benefits Research

1. produces an accurate and explainable DHF diagnosis model;
2. enriches the literature on the use of ML in Indonesian clinical data;
3. supports the development of a Clinical Decision Support System (CDSS) for primary health care services

2. Research Methods

This study uses a quantitative approach based on machine learning with the Random Forest algorithm as the main method for classifying patient clinical symptom data. The analysis procedure was carried out systematically through the stages of modeling, performance evaluation, and statistical validation to ensure that the predictions produced were not only numerically accurate but also had clinical relevance in supporting the diagnosis of dengue hemorrhagic fever (DHF).

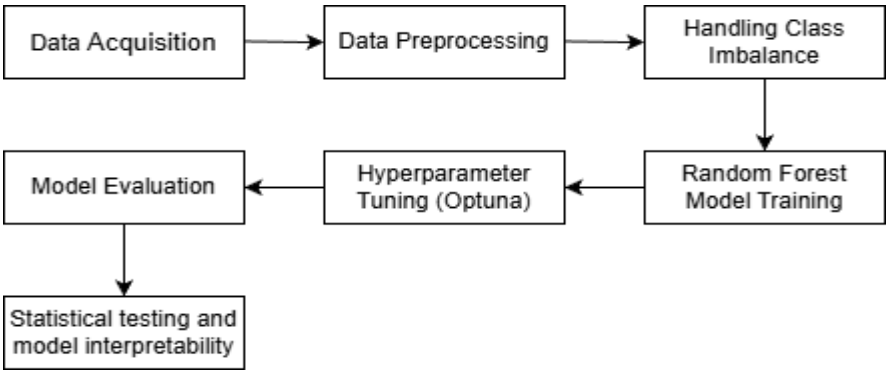

Fig. 1. Research Stages

2.1. Data Acquisition

The research dataset was obtained from the electronic medical records of patients at the Kosasih Group Clinic who had a history of visits with suspected dengue fever. The data is available in a structured format (CSV/Excel) and contains clinical symptom variables as predictors, such as headache, dizziness, nausea, vomiting, sore throat, joint pain, skin rash, and bleeding gums, while the diagnosis of dengue fever is used as a classification label. The acquisition process was carried out through an Extract–Transform–Load (ETL) flow, which began with data extraction based on a specific time range and clinical criteria [13]. The transformation stage included the harmonization of medical terms, the encoding of features into numerical values, and the standardization of data formats. Next, the Load stage is carried out by transferring the processed data into research storage for analysis purposes. All data has undergone de-identification and security procedures in accordance with medical research ethics standards. Thus, the acquisition process produces a clinical symptom dataset that is ready for use in the modeling stage.

2.2. Data Preprocessing

Pre-processing is performed to ensure optimal data quality prior to model training. Duplicate data are identified and handled through a probabilistic record linkage approach so that only unique records are used. Missing data mechanisms are analyzed to determine the type of missingness (MCAR, MAR, MNAR), then imputed using model-based methods such as missForest, which has been proven effective for non-linear clinical data [14]. Categorical variables are encoded using regularized target encoding, while numeric variables are normalized using z-scores or min–max scaling to maintain scale consistency. All of these steps ensure that the dataset is clean, free of structural bias, and representative of actual clinical conditions.

2.3. Handling Class Imbalance

The low proportion of positive dengue cases causes class imbalance, which has the potential to reduce the sensitivity of the model. To overcome this condition, this study applies a combinative strategy at both the data and algorithmic levels. At the algorithmic level, class weighting is used to give higher weight to prediction errors in minority classes. Meanwhile, at the data level, moderate SMOTE techniques are applied, including variants such as Borderline-SMOTE or SMOTE-ENN, to improve the representation of the minority class while limiting the emergence of noise. This hybrid approach is designed to achieve a balance between model sensitivity and stability in the context of actual clinical application [15].

2.4. Random Forest Model Training

The Random Forest model is trained using a dataset that has undergone all stages of pre-processing. Relevant features are selected through variable screening to reduce model variance and improve modeling efficiency. Key parameters, such as the number of trees, maximum depth, and minimum node size, are carefully set to prevent overfitting, especially in datasets with limited size. The training process also integrates inverse-probability weighting to strengthen the representation of minority classes in model learning. The entire training sequence is applied in a structured manner to obtain a model that is stable, accurate, and suitable for application in a clinical context [16].

2.5. Hyperparameter Tuning (Optuna)

Hyperparameter optimization was performed using the Bayesian Optimization approach through Optuna. The search space included core parameters in Random Forest, namely n_estimators, max_depth, min_samples_leaf, and max_features. Optuna utilizes the Tree-Structured Parzen Estimator (TPE) algorithm to identify the most promising parameter combinations and applies a pruning mechanism to stop inefficient iterations [17]. The tuning process was run for 50 trials, and the best configuration was validated using nested cross-validation to prevent information leakage. This approach ensures that the model obtains optimal settings without reducing interpretability.

2.6. Model Evaluation

Model performance evaluation was conducted using discrimination and calibration metrics. AUROC was used to assess the model's ability to distinguish between DHF and non-DHF cases, while AUPRC was chosen because it is more representative of data with a low proportion of minority classes. Calibration quality is measured using the Brier Score to assess the suitability between the predicted probability and the actual value [18]. Additional evaluations include calibration plots and confusion matrices to provide a more comprehensive interpretation. All metrics are assessed using nested cross-validation so that the performance obtained is free from bias and has good generalization capabilities.

2.7. Statistical testing and model interpretability

The final stage of the study involved statistical testing and model interpretability analysis using SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-Agnostic Explanations). SHAP analysis was used to estimate the contribution of each feature to the model prediction, both at the global and individual levels, as well as to assess the consistency of the influence between symptoms. Visualizations in the form of beeswarm plots, bar plots, and dependence plots are applied to illustrate the level of feature importance and clinical interaction patterns [19]. Meanwhile, LIME is used to generate local interpretations in individual cases to explain the basis for the model's decision-making in specific patients so that the prediction results are easier for medical personnel to understand. The combination of SHAP and LIME provides scientific justification for model decisions and strengthens the transparency and accountability of prediction systems in a clinical context.

3. Results and Discussion

3.1. Data Acquisition

The research dataset was obtained from anonymized patient clinical records and compiled in digital format (.xlsx) with a total of 1,200 observations and 11 variables. Each entry represents one patient, while the recorded features include binary clinical symptoms (fever, headache, nausea, vomiting, muscle pain, joint pain, skin rash, and gum bleeding), one categorical feature (gender), and one target label (DHF diagnosis). All data were processed using Python in the Google Colab environment. The dataset does not contain any personal identity information, thus meeting data ethics and privacy standards.

```
Reading: /content/drive/MyDrive/Skripsi/DBD 22.xlsx
Dataset shape: (1200, 11)
```

	No	Jenis Kelamin	Demam	Sakit Kepala	Mual	Muntah	Nyeri Otot	Nyeri Sendi	Ruam Kulit	Gusi Berdarah	Diagnosis DBD
0	1	Perempuan	1	0	1	0	1	0	0	0	1
1	2	Laki-laki	1	0	0	1	0	0	0	0	1
2	3	Perempuan	1	1	1	0	1	1	1	1	1
3	4	Perempuan	0	1	0	0	1	1	1	1	1
4	5	Perempuan	1	0	1	0	0	0	0	0	1

Fig. 2. Initial Data Snapshot

3.2. Data Preprocessing

The pre-processing step includes checking for missing values, data types, duplicates, and setting target columns. The analysis results show that there are no missing values and no duplicates. The categorical feature "Gender" is encoded numerically using label encoding, while other symptom features are retained as binary. Since most features are 0/1 variables, additional normalization is not necessary. The cleaned dataset is then used for class balancing.

```
Target column: Diagnosis DBD
Missingness fraction (top 30):
No                0.0
Jenis Kelamin     0.0
Demam             0.0
Sakit Kepala      0.0
Mual              0.0
Muntah            0.0
Nyeri Otot        0.0
Nyeri Sendi       0.0
Ruam Kulit        0.0
Gusi Berdarah     0.0
Diagnosis DBD     0.0
dtype: float64
Numeric cols: 8 Categorical cols: 1
```

Fig. 3. Results of Missing Data and Column Type Checks

3.3. Handling Class Imbalance

The initial distribution showed an imbalance between the DBD and non-DBD classes. The SMOTE (Synthetic Minority Over-sampling Technique) method was applied to generate synthetic samples in the minority class. After the resampling process, both classes became balanced (486:486). This balancing is important to improve the sensitivity of the model in detecting DBD cases and prevent bias towards the majority class.

```
After SMOTE: Diagnosis DBD
0    486
1    486
Name: count, dtype: int64
```

Fig. 4. Class Distribution After SMOTE

3.4. Random Forest Model Training

The Random Forest Classifier model was trained using the SMOTE dataset, with default parameters using automatic balancing (class_weight='balanced') and reproducibility through random_state=42. Initial evaluation with cross-validation resulted in an ROC-AUC of 0.9533, indicating excellent discriminatory ability. This performance became the basis for continuing hyperparameter tuning.

```
CV ROC-AUC (RF on resampled train): 0.9533037191920025
RandomForestClassifier
RandomForestClassifier(class_weight='balanced', n_jobs=-1, random_state=42)
```

Fig. 5. Initial ROC-AUC Values of the Random Forest Model Hyperparameter Tuning

3.5. Hyperparameter Tuning (Optuna)

Hyperparameter optimization was performed using Optuna, focusing on four main parameters: n_estimators, max_depth, max_features, and min_samples_leaf. From more than 20 experiments, the best configuration was obtained:

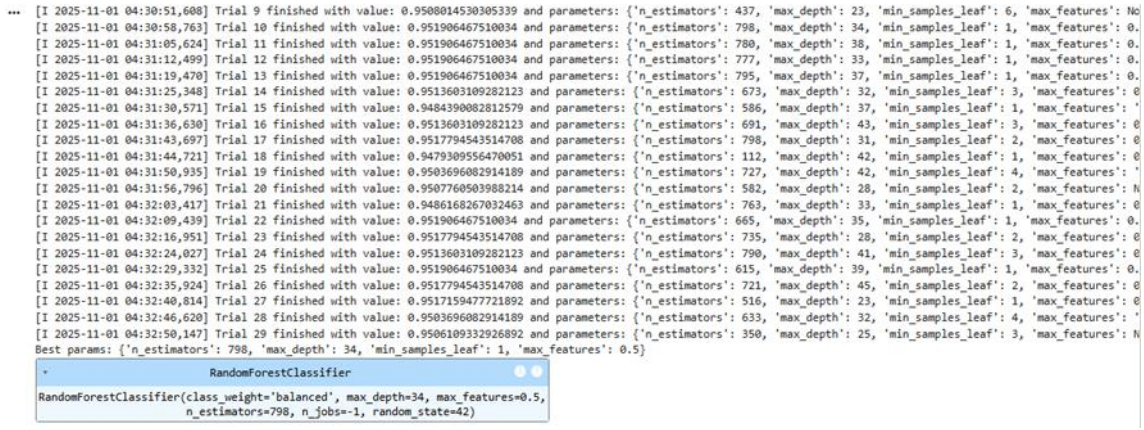

Fig. 6. Hyperparameter Tuning Results with Optuna

This parameter produces a model that balances complexity and generalization well. Hyperparameter optimization was performed using Optuna, focusing on four main parameters: n_estimators, max_depth, max_features, and min_samples_leaf.

From more than 20 experiments, the best configuration was obtained:

Table 1. Hyperparameter Tuning

Parameter	Value
n_estimators	798
max_depth	34
max_features	0.5
min_samples_leaf	1

3.6. Model Evaluation

The evaluation results on the test data show the following performance: The model has excellent discrimination with very good ROC and PR curves. The calibration plot shows accurate prediction probabilities.

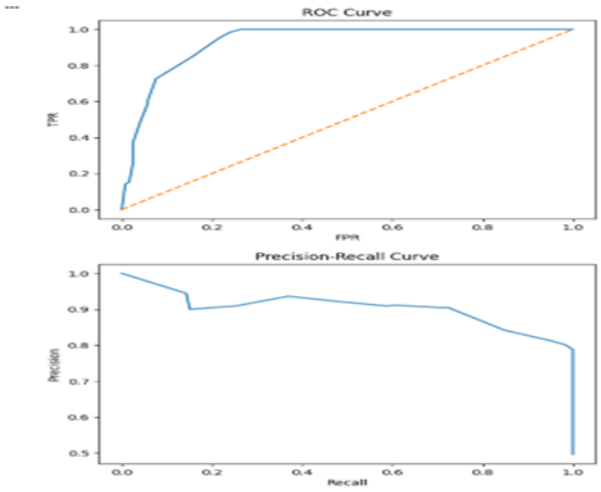

Fig 7. ROC Curve and Precision–Recall Curve

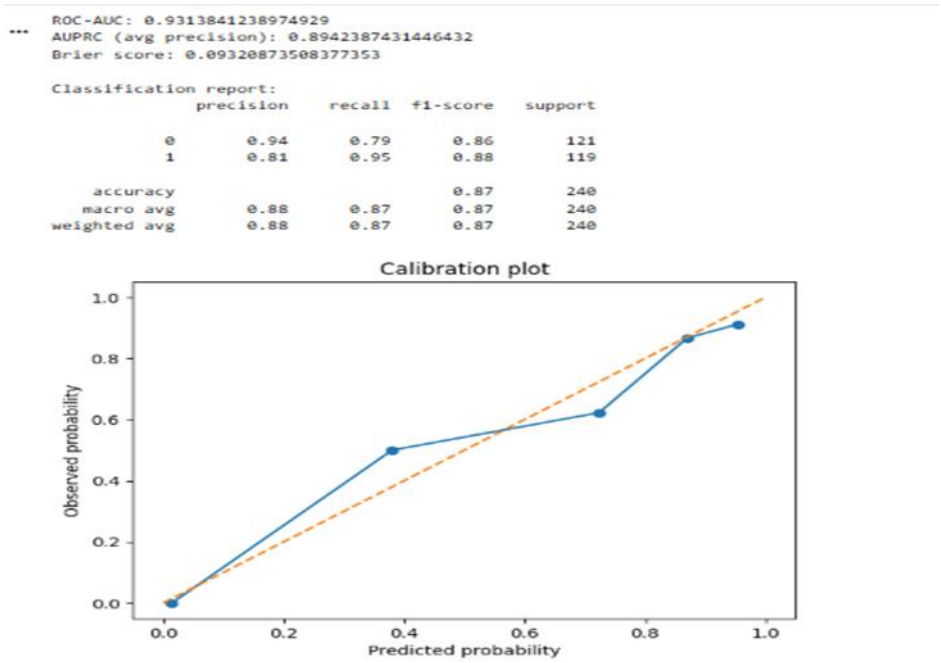

Fig. 8. Random Forest Model Calibration Plot

The confusion matrix shows that the model has high sensitivity to dengue cases and a low false negative rate, making it safe for clinical use. Statistical Testing

AUROC testing was conducted to measure the stability of the model's performance. The results:

1. **Mean AUROC:** 0.9324
2. **95% Confidence Interval:** (0.9050 – 0.9581)

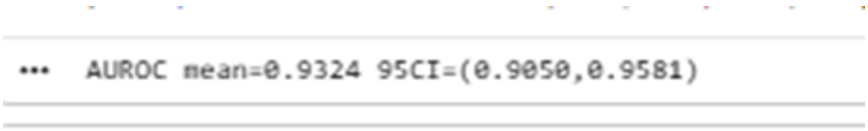

Fig 9. AUROC and Confidence Interval

A narrow CI indicates that the model is stable across various data subsets.

3.7. Statistical testing and model interpretability

SHAP analysis shows the most influential features: vomiting, nausea, fever, skin rash, and bleeding gums. SHAP visualization confirms the global pattern of feature contribution, while LIME provides local explanations for each patient prediction.

Fig. 10. SHAP Summary Plot

This interpretation helps increase clinical confidence in the model because it is consistent with the epidemiological pattern of dengue hemorrhagic fever.

3.8 Model Visualization and Summary

Visual analysis includes feature importance, ROC and PR curves, confusion matrix, and decision curve analysis.

Key findings:

1. Vomiting, nausea, and fever are the most dominant predictors.
2. The confusion matrix shows high sensitivity to the DBD class and low FN.
3. Decision Curve Analysis shows a higher net clinical benefit compared to conventional approaches.

The model obtains:

Table 2. Metrix Evaluation

Metrix Evaluation	Value
Akurate	**0.8667**
F1-Score	**0.8760**
ROC-AUC	**0.9314**

The model meets the criteria for an excellent classifier for dengue diagnosis.

4. Conclusion

This study shows that the Random Forest algorithm can improve the accuracy of early diagnosis of dengue hemorrhagic fever (DHF) based on clinical symptoms in primary health care services. The developed model produces high predictive performance (AUROC 0.9314; accuracy 0.88; AUPRC 0.8942), making it effective in distinguishing DHF from other febrile illnesses. The use of medical record data through the CRISP-DM approach and the integration of interpretability methods such as SHAP and LIME produced a model that is not only accurate but also transparent and clinically acceptable. These findings confirm the model's potential for use as a clinical decision support system without replacing the role of doctors in the diagnosis process. However, this study still requires external validation at other health facilities to ensure the generalization of the model. Prospective implementation in real clinical practice is also important to assess the model's impact on diagnostic effectiveness. Developing a digital application that integrates predictive models and adding simple laboratory variables can improve the model's sensitivity and usefulness. Further research is also recommended to include longitudinal analysis to develop disease progression predictions. Overall, the results of this study can serve as a basis for the development of an artificial intelligence-based Clinical Decision Support System in primary health care facilities.

References

[1] G. S. Collins *et al.*, "TRIPOD+AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods," *BMJ*, vol. 385, p. e078378, 2024.

[2] N. Sharif, M. R. Islam, and M. Hasan, "Evolving epidemiology, clinical features, and genotyping of dengue outbreaks in Bangladesh, 2000–2024: A systematic review," *Front. Microbiol.*, vol. 15, p. 1481418, 2024.

[3] H. Long *et al.*, "Annual global dengue dynamics are related to multi-source factors revealed by a machine learning prediction analysis," *PLoS Negl. Trop. Dis.*, vol. 19, no. 6, p. e0013232, 2025.

[4] H. Xia and X. Dong, "The global, regional, and national burden trends of dengue among adults aged 20–49 from 1990 to 2021," *Sci. Rep.*, vol. 15, p. 26761, 2025.

[5] J. Jung, J. Dai, B. Liu, and Q. Wu, "Artificial intelligence in fracture detection with different image modalities and data types: A systematic review and meta-analysis," *PLOS Digit. Heal.*, vol. 3, no. 1, p. e0000438, 2024.

[6] R. T. Subarna and Z. Al Saiyan, "Understanding the unprecedented 2023 dengue outbreak in Bangladesh: A data-driven analysis," *IJID Reg.*, vol. 12, p. 100406, 2024.

[7] X. Y. Leung *et al.*, "A systematic review of dengue outbreak prediction models: Current scenario and future directions," *PLoS Negl. Trop. Dis.*, vol. 17, no. 2, p. e0010631, 2023.

[8] M. Y. Ng *et al.*, "Perceptions of dataset experts on important characteristics of health datasets ready for machine learning: A qualitative study," *JAMA Netw. Open*, vol. 6, no. 2, p. e2812417, 2023.

[9] S. R. da Silva Neto, T. Tabosa Oliveira, I. V Teixeira, S. B. Aguiar de Oliveira, V. Souza Sampaio, and T. Lynn, "Machine learning and deep learning techniques to support clinical diagnosis of arboviral diseases: A systematic review," *PLoS Negl. Trop. Dis.*, vol. 16, no. 1, p. e0010061, 2022.

[10] B. C. Bohm *et al.*, "Utilization of machine learning for dengue case screening," *BMC Public Health*, vol. 24, 2024.

[11] G. Gupta *et al.*, "DDPM: A dengue disease prediction and diagnosis model using sentiment analysis and machine learning algorithms," *Diagnostics*, vol. 13, no. 6, p. 1093, 2023.

[12] R. Zargari Marandi, P. Leung, C. Sigera, D. D. Murray, P. Weeratunga, and D. Fernando, "Development of a machine learning model for early prediction of plasma leakage in suspected dengue patients," *PLoS Negl. Trop. Dis.*, vol. 17, no. 3, p. e0010758, 2023.

[13] A. Lamer, C. Saint-Dizier, N. Paris, and E. Chazard, "Data lake, data warehouse, data mart, and feature store: Their contributions to the complete data reuse pipeline," *JMIR Med. Informatics*, vol. 12, p. e54590, 2024.

[14] S. Islam Khan and A. S. M. L. Hoque, "SICE: an improved missing data imputation technique," *J. Big Data*, vol. 7, p. 37, 2020.

[15] R. Hassanzadeh, M. Farhadian, and H. Rafieemehr, "Hospital mortality prediction in traumatic injuries patients: Comparing different SMOTE-based machine learning algorithms," *BMC Med. Res. Methodol.*, vol. 23, p. 101, 2023.

[16] P. Studi and S. Informasi, "KLASIFIKASI EMOSI PADA TWEET BERBAHASA," no. 86.

[17] L. Barreñada, P. Dhiman, D. Timmerman, A.-L. Boulesteix, and B. Van Calster, "Understanding overfitting in random forest for probability estimation: A visualization and simulation study," *Diagnostic Progn. Res.*, vol. 8, p. 14, 2024.

[18] R. D. Riley, "Evaluation of clinical prediction models (Part 2): Calibration, discrimination and overall performance metrics," *BMJ*, vol. 384, 2024.

[19] H. Wang, Q. Liang, J. T. Hancock, and T. M. Khoshgoftaar, "Feature selection strategies: A comparative analysis of SHAP-value and importance-based methods," *J. Big Data*, vol. 11, p. 44, 2024.