BIMA
BULLETIN OF INTELLIGENT MACHINES AND ALGORITHMS
Vol. 1 No.2 January 2026
E-ISSN: 3132-5115
DOI: 10.65780/bima.v1i2.10

Improving Dengue Case Prediction using Feature Engineering and SHAP in Bandung City on Climate Demographic Data

Taufik Abdul Aziz[1*], Iqbal Ismayadi[2]

Informatics[1,2]
Universitas Informatika dan Bisnis Indonesia, Bandung, Indonesia[1,2]
https://unibi.ac.id/[1]
taufik.aa22@student.unibi.ac.id[1*]

Abstract. Dengue Hemorrhagic Fever remains a major public health concern in urban areas of Indonesia, particularly in Bandung City, due to its fluctuating incidence and strong dependence on environmental and population factors. This study focuses on improving dengue case prediction by integrating climate and demographic data through systematic feature engineering and explainable machine learning based on the Random Forest algorithm. Historical dengue case data from Bandung City were used to develop and evaluate the proposed prediction model. The evaluation results show that the Random Forest model achieved an R^2 value of 0.9032 and an RMSE of 2.3748, indicating reliable predictive performance and good generalization capability. The applied feature engineering strategy effectively enhanced data representation by capturing temporal dynamics, case growth patterns, and interactions among climate variables. Furthermore, model interpretability was improved through the application of Explainable Artificial Intelligence using SHAP, which revealed that temporal features derived from previous dengue case trends were the most influential factors, followed by climate interaction variables. These findings demonstrate that the proposed approach improves prediction accuracy while providing transparent and epidemiologically meaningful insights to support data driven dengue early warning systems at the regional level.

Key words: Dengue Fever; Random Forest; Feature Engineering; Explainable AI; Bandung City

1. INTRODUCTION

Dengue hemorrhagic fever (DHF) is an infectious disease transmitted by the Aedes aegypti and Aedes albopictus mosquito vectors, which to date remains a major public health problem in various tropical and subtropical countries, including Indonesia. Specifically in the city of Bandung, the accumulation of cases reaching 7,310 in 2024 and decreasing significantly to around 3,000 cases in 2025 requires continued vigilance, along with predictions of a resurgence cycle expected to occur in 2026 [1], [2]. The disease is transmitted by the Aedes aegypti and Aedes albopictus mosquito vectors, with incidence rates that tend to fluctuate and are influenced by various environmental factors and population characteristics [3]. The World Health Organization notes that dengue incidence has continued to increase globally in recent decades, causing significant social, economic, and health burdens [4]. This condition requires an accurate and reliable dengue case prediction system to support preventive and responsive public health intervention planning.

In the context of dengue epidemiology, climatic factors such as temperature, rainfall, and humidity are known to play an important role in influencing the life cycle of mosquito vectors and the dynamics of dengue virus transmission [5]. In addition, demographic factors such as population density, age structure, and urbanization rates also contribute to the risk of disease spread [6]. However, the relationship between these variables is complex, nonlinear, and interactive, making it difficult to model effectively using conventional statistical approaches. Therefore, the use of Machine Learning (ML) methods is a promising alternative for capturing hidden patterns in multidimensional and large-scale data.

However, the application of Machine Learning models in dengue case prediction still faces a number of challenges. One of the main challenges is the quality and representation of features used in the modeling process. Many previous studies still utilize climate or demographic data separately, with feature

Vol. 1 No.2 January 2026
E-ISSN: 3132-5115
DOI: 10.65780/bima.v1i2.10

representation used directly without adequate feature engineering, so that changes in data patterns over time and the relationship between variables have not been optimally utilized [7], [8]. In addition, highly complex Machine Learning models often have low interpretability, limiting their use in public health decision-making contexts that require clear explanations [9]. This lack of transparency can hinder trust and adoption of prediction models in real decision-making.

In response to these challenges, this study aims to improve dengue case prediction performance through the systematic application of feature engineering to climate and demographic data, as well as integrating the Explainable Artificial Intelligence (XAI) approach to improve model interpretability. Feature engineering is performed to extract and construct informative features, such as temporal lag variables, statistical aggregations, and nonlinear transformations, which are expected to represent environmental and population dynamics more accurately [10]. Meanwhile, XAI is used to reveal the relative contribution of each feature to the model prediction, so that the relationship between climate, demographic factors, and the increase in dengue cases can be explained transparently and logically [11].

This study aims to develop a dengue case prediction model that has high performance and is able to provide a clear understanding of the factors that influence prediction results. Specifically, this study evaluates the effect of feature engineering on improving Machine Learning model performance, as well as identifying the relative contribution of climate and demographic variables through the Explainable Artificial Intelligence (XAI) approach. The integration of comprehensive feature engineering and XAI-based analysis is the main contribution of this study, which distinguishes it from previous studies that generally focus solely on improving accuracy. Through a more in-depth analysis of the dynamics of environmental and demographic factors, the results of this study are expected to not only enrich the study of dengue epidemiological prediction, but also provide a strong scientific basis for the development of more effective and data-driven early warning systems.

2. RELATED WORK

Various studies have utilized machine learning to predict dengue cases by incorporating meteorological and demographic variables as model inputs. In a study the application of XGBoost to dengue meteorological data in Singapore showed that this model achieved an R^2 value of 0.83, MAE 89.12, and RMSE 156.07 when predicting dengue cases, highlighting the ability of boosting algorithms to capture non-linear relationships between complex climate variables [12].

In addition, research in recent years has shown that Random Forest and ensemble learning algorithms remain competitive approaches in dengue case prediction based on climate and epidemiological data. For example, dengue prediction models built with Random Forest and XGBoost on environmental and weather datasets show that XGBoost provides better prediction performance with lower error values than Random Forest, confirming the ability of ensemble learning to capture nonlinear variations in infectious disease data [13]. Meanwhile, several local studies utilizing classification methods on demographic and clinical data show that Random Forest can achieve an accuracy of up to 90.0% and an AUC of 0.967 in the task of early detection of dengue hemorrhagic fever cases, while maintaining prediction stability through cross-validation [14].

Other studies focusing on dengue modeling in various regions also show variations in model performance depending on the methodology and dataset used. A study in the coastal region of Sumatra using Random Forest and SVM reported that the Random Forest model had a lower MSE than SVM, although error metrics such as RMSE or full accuracy were not reported [15].

In recent years, explainable AI approaches such as SHAP have begun to be used to improve the interpretability of dengue prediction models by revealing the contribution of climatic and demographic variables to prediction results. However, the application of systematic feature engineering and SHAP analysis on an urban scale is still limited, especially in cities with complex characteristics such as Bandung. This study aims to fill this gap by combining feature engineering and SHAP to produce an accurate and interpretable dengue prediction model.

3. METHODS

This study proposes a methodological framework for dengue case prediction that integrates climate and demographic data through systematic data processing, feature engineering, machine learning modeling, and explainable AI approaches. The proposed method flow is designed not only to improve prediction accuracy but also to provide clear interpretations of each feature's contribution to influencing prediction results. Figure 1 presents an overview of the proposed method stages used in this study.

BIMA
BULLETIN OF INTELLIGENT MACHINES AND ALGORITHMS
Vol. 1 No.2 January 2026
E-ISSN: 3132-5115
DOI: 10.65780/bima.v1i2.10

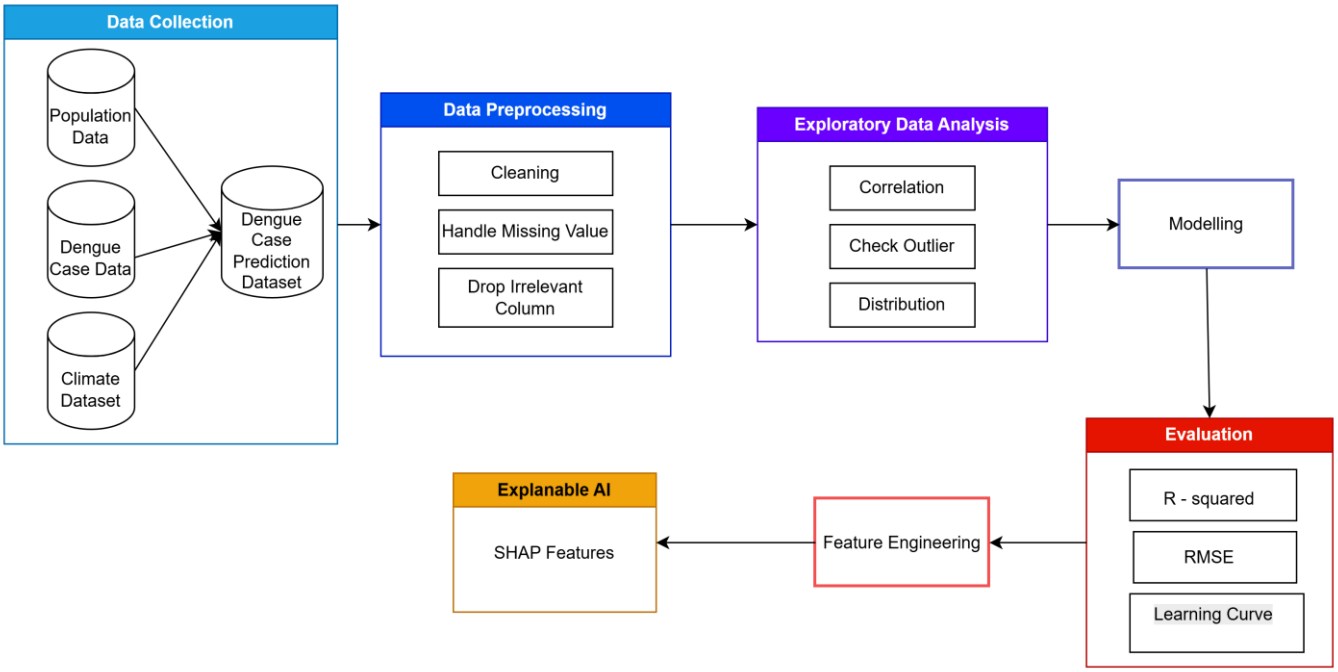

Fig. 1. Flow Study

The approach utilized in this research is organized into multiple essential phases to guarantee a systematic and repeatable evaluation process. Every phase is intended to meet particular goals, encompassing data preparation, feature creation, predictive modeling, and performance assessment. The comprehensive account of every methodological phase is outlined below

3.1. Data Collection

This study uses three main sources of data that are integrated with each other, namely dengue fever case data, climate data, and population density data. Dengue fever case data was obtained from regional health agencies and includes the number of cases per subdistrict in Bandung City on an annual basis. Climate data includes variables such as average temperature, average humidity, and rainfall obtained from the Meteorology, Climatology, and Geophysics Agency (BMKG). Meanwhile, population density data per subdistrict is obtained from the Central Statistics Agency (BPS). The three datasets are then combined based on geographical (subdistrict) and time period (year) compatibility to form a DHF case prediction dataset.

3.2. Data Preprocessing

Preprocessing steps are carried out to ensure data quality and consistency before further analysis. This process includes data cleaning by removing duplicate data and irrelevant columns, as well as handling missing values using an appropriate imputation approach. In addition, categorical columns such as subdistrict names are encoded using one-hot encoding so that they can be processed by machine learning algorithms. Numeric feature normalization is also applied to equalize the scale between variables, especially in models that are sensitive to data scale.

3.3. Exploratory Data Analysis

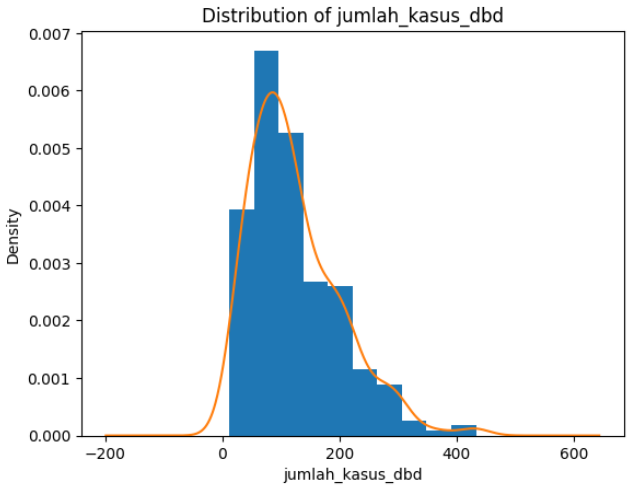

Fig. 2. Distribution of dengue

Figure 2 shows a right-skewed pattern, where most subdistricts have low to moderate numbers of cases, while a few subdistricts experience very high numbers of cases. This indicates an imbalance in the distribution of dengue fever cases between subdistricts, which may be influenced by environmental and demographic factors.

BIMA
BULLETIN OF INTELLIGENT MACHINES AND ALGORITHMS
Vol. 1 No.2 January 2026
E-ISSN: 3132-5115
DOI: 10.65780/bima.v1i2.10

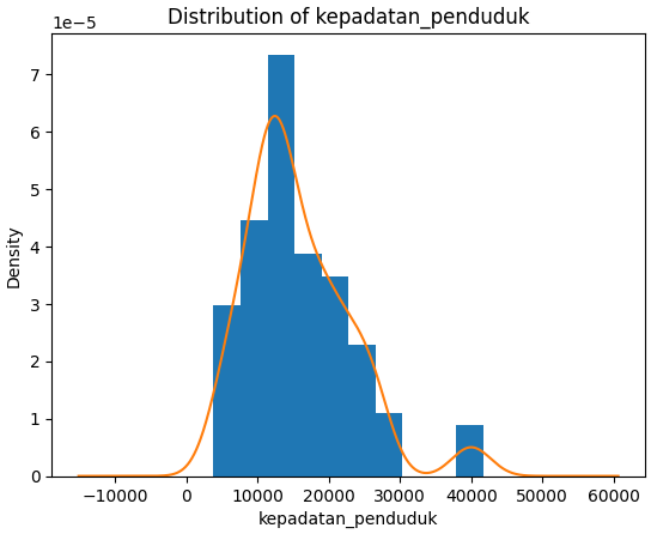

Fig. 3. Distribution of population density

Figure 3 also asymmetrical and tends to be skewed to the right, with the majority of subdistricts having medium density levels. Several subdistricts with very high density have the potential to become areas at risk of an increase in dengue fever cases due to greater human interaction.

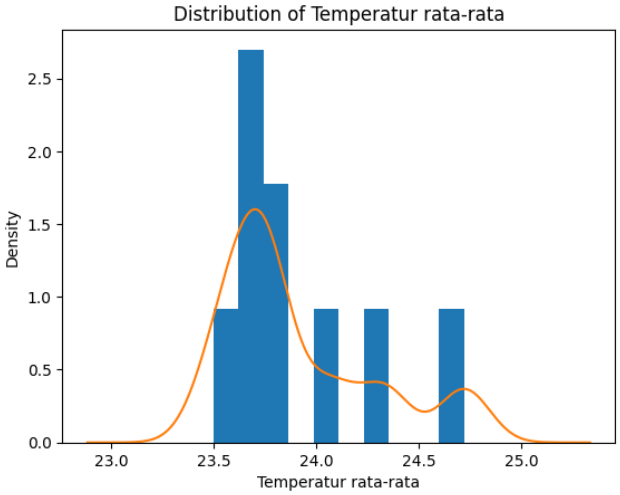

Fig. 4. Distribution of average temperature

Figure 4 is relatively narrow and concentrated, indicating that temperature variations between subdistricts and years are not particularly large. Nevertheless, these small differences in temperature range are still relevant because they can affect the life cycle of mosquito vectors that cause dengue fever.

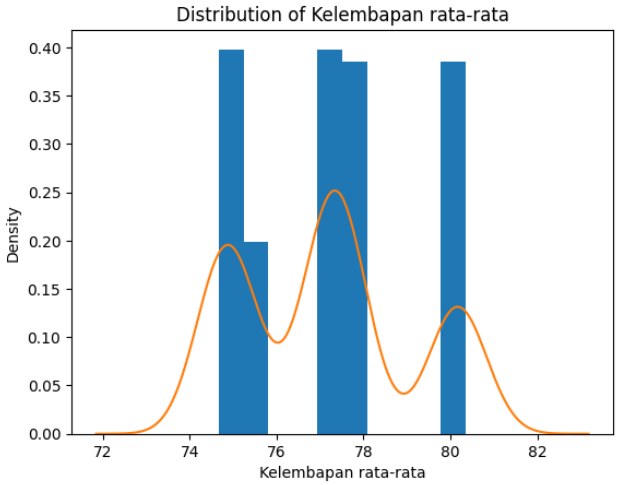

Fig. 5. Distribution of average humidity

Figure 5 shows a multimodal pattern with concentrations of values in the medium to high humidity range. This condition indicates that variations in humidity levels between regions and over time have the potential to play an important role in creating an environment that supports the development of mosquito vectors that cause dengue fever.

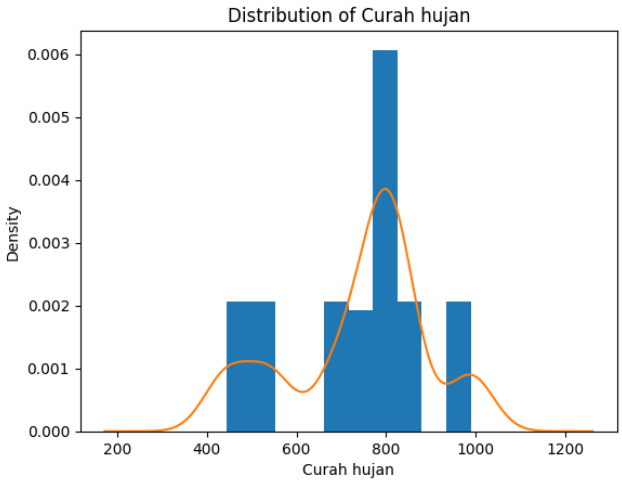

Fig. 6. Distribution of Rainfall

Figure 6 shows a multimodal pattern, indicating variations in rainfall conditions between periods and regions. High rainfall values have the potential to create an environment conducive to mosquito breeding, making them a relevant predictor of dengue fever cases.

3.4. Feature Engineering

In order for the model to capture more complex patterns, we perform feature engineering. This process is done by adding new relevant features. These interaction features are created from combinations of climate variables, such as the interaction between rainfall and humidity, as well as temperature and humidity. Then, we apply non-linear transformations, such as squaring the temperature values and

Vol. 1 No.2 January 2026
E-ISSN: 3132-5115
DOI: 10.65780/bima.v1i2.10

performing logarithmic transformations on population density. In addition, we also add epidemiological ratios, namely the number of dengue fever cases per population, to show the relative risk. Temporal features are also constructed using the lag of dengue fever cases in the previous year and the rolling mean to capture temporal patterns between years.

3.5. Modeling

The dataset was divided into training and testing sets, where 80% of the data were utilized for model training and the remaining 20% were reserved for model evaluation. This strategy was adopted to assess the generalization capability of the models on unseen data. To further ensure the robustness and stability of the predictive performance, k-fold cross-validation was applied during the training phase.

Random Forest regression is used as the first ensemble learning model because of its robustness in modeling nonlinear relationships and feature interactions that are often found in epidemiological data. Random Forest is a bagging-based ensemble method that builds decision trees independently using bootstrap samples and random feature selection. The final prediction is obtained by aggregating the predictions from all individual trees, which effectively reduces variance and reduces overfitting [16]. Mathematically, the Random Forest regression prediction is defined as Eq (1).

$$\hat{y}(x) = (1/T)\sum(t=1)^{T} f_t(x) \quad (1)$$

where T denotes the total number of decision tree and $f_t(x)$ represents the prediction of the $(t-th)$ decision tree for input x.

Extreme Gradient Boosting (XGBoost) is applied as a boosting-based ensemble model to improve prediction accuracy. XGBoost builds decision trees sequentially, where each new tree is trained to minimize the residual error of the previous ensemble. Unlike Random Forest, XGBoost optimizes a regularized objective function that balances prediction accuracy and model complexity [17]. The XGBoost objective function is formulated as Eq (2).

$$L = \sum(i=1)^{n} l(y_i, \hat{y}_i) + \sum(k=1)^{K} \Omega(f_k) \quad (2)$$

where $l(y_i, \hat{y}_i)$ is the loss function (squared error for regression), f_k represents the$(k = th)$ decision tree $\Omega(f_k)$is the regularization term defined as Eq (3).

$$\Omega(f) = \gamma T + (1/2)\lambda \sum(j=1)^{T} w_j^2 \quad (3)$$

with T denoting the number leaves in the tree, w_j representing the leaf weights, and γ and λ acting as regularization parameters to control model complexity

By employing both Random Forest and XGBoost, this study provides a comparative analysis between bagging-based and boosting-based ensemble approaches, enabling a comprehensive evaluation of their effectiveness in predicting dengue cases using climate and demographic features.

3.6. Evaluation

The performance of the proposed models was evaluated using the coefficient of determination (R^2) and the Root Mean Squared Error (RMSE). The R^2 metric measures the proportion of variance in the actual dengue case data that can be explained by the model predictions and is defined as Eq (4).

$$R^2 = 1 - (\sum(i=1)^{n}(y_i - \hat{y}_i)^2)/(\sum(i=1)^{n}(y_i - \bar{y})^2) \quad (4)$$

where y_i represents the actual value, y_i denotes the predicted value, and $\bar{y}$ is the mean of the observed values. In addition, RMSE was used to quantify the average magnitude of prediction errors in the original scale of the target variable, which is formulated as

3.7. Explainable AI

To improve model interpretability, the Explainable AI approach was applied using SHAP (SHapley Additive exPlanations). SHAP analysis was used to identify the contribution of each feature to the prediction of DHF cases, thereby providing a more transparent understanding of the most influential factors in the prediction model.

4. RESULTS AND DISCUSSIONS

4.1. Results

To assess the performance of the Dengue Hemorrhagic Fever (DHF) case prediction model, this study conducted a comparative evaluation of two machine learning algorithms, namely Random Forest and XGBoost. The evaluation was carried out using the R-square (R^2) and Root Mean Square Error (RMSE) metrics to provide a comprehensive overview of the model's ability to explain data variability and the level of prediction error against actual values. The use of these two metrics allows for a more objective analysis of model performance, both in terms of the accuracy of data pattern representation and the numerical accuracy of prediction results. The results of the comparison of the performance of the two models are summarized in Table 1.

TABLE 1. Compared evaluation model of machine learning.

Model	Evaluation	
	R^2	RMSE
Random Forest	0.9032	2.3748
XGBoost	0.9520	1.6663

Vol. 1 No.2 January 2026
E-ISSN: 3132-5115
DOI: 10.65780/bima.v1i2.10

Based on the evaluation results shown in Table 1, the XGBoost model produced a higher R-square value of 0.9520 compared to Random Forest, which obtained a value of 0.9032. This finding indicates that XGBoost has a stronger ability to explain the variability of Dengue Hemorrhagic Fever (DHF) case data, so that the relationship pattern between input variables and the number of cases can be represented more complexly. In addition, the lower RMSE value in XGBoost, which is 1.6663, indicates a smaller prediction error rate compared to Random Forest with an RMSE of 2.3748. Numerically, these results suggest that XGBoost provides predictions that are closer to the actual values. However, this difference in performance needs to be analyzed further by considering the learning characteristics of the model and its potential for generalization, given that the high evaluation value at this stage does not fully reflect the stability of the model when faced with different data.

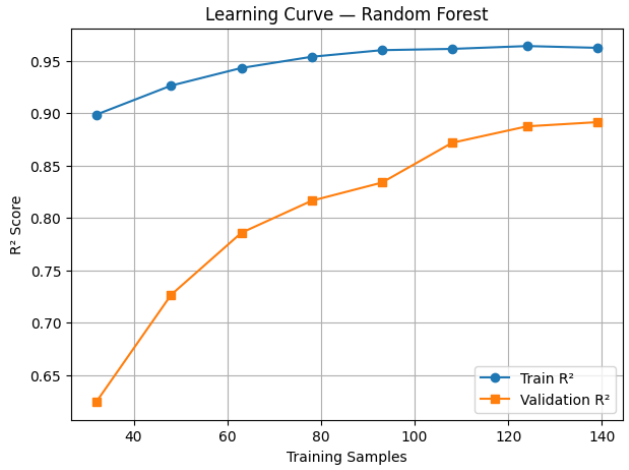

Fig. 7. Learning Curve of Random Forest

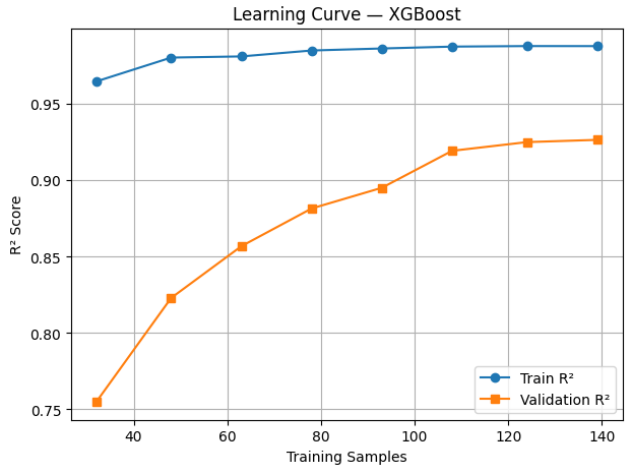

Fig. 8. Learning Curve of XGBoost

The learning curve analysis in both figures shows a clear difference in learning characteristics between the XGBoost and Random Forest models. In XGBoost, the training data R-square curve shows a very high value even with a relatively small amount of training data and tends to stabilize as the number of samples increases. Meanwhile, the validation curve experiences a gradual increase and only approaches the training curve when the training data size becomes larger. This pattern indicates that XGBoost has high model capacity and is able to capture complex patterns quickly, but in the early stages of learning, there is still a gap between training and validation performance, reflecting a potential tendency toward overfitting.

Unlike XGBoost, Random Forest exhibits a more gradual learning pattern and a better balance between the training and validation curves. Although the R-square value on the training data is slightly lower than that of XGBoost, the improvement in performance on the validation data is consistent as the number of samples increases. The gap between the training and validation curves in Random Forest is relatively smaller, especially with medium to large training data sizes, indicating a more stable model generalization ability. This characteristic is important in the context of dengue case prediction, which is influenced by temporal variability and dynamic environmental factors.

Overall, the learning curve results confirm that although XGBoost excels in terms of numerical performance, Random Forest exhibits more robust and reliable learning behavior in response to changes in data volume. This finding reinforces the basis for selecting Random Forest as a more suitable model for further analysis, particularly in the context of interpretability using the Explainable Artificial Intelligence (XAI) approach, where model stability and consistency are crucial aspects.

BIMA
BULLETIN OF INTELLIGENT MACHINES AND ALGORITHMS
Vol. 1 No.2 January 2026
E-ISSN: 3132-5115
DOI: 10.65780/bima.v1i2.10

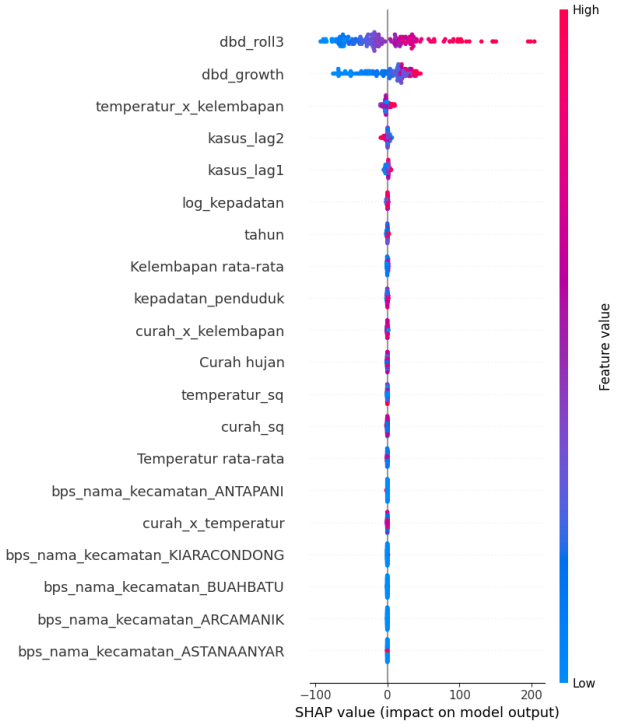

Fig. 9. SHAP beeswarm

Figure 9 shows that temporal-based variables and previous case dynamics, such as DHF_roll3 and DHF_growth, have the most dominant influence on model predictions. Positive SHAP values on these features indicate that an increase in case trends in the previous period contributes significantly to an increase in DHF case predictions in the following period. This finding confirms that historical case patterns are the main determinants in the formation of predictions, in line with the characteristics of DHF spread, which is seasonal and has a strong temporal dependence.

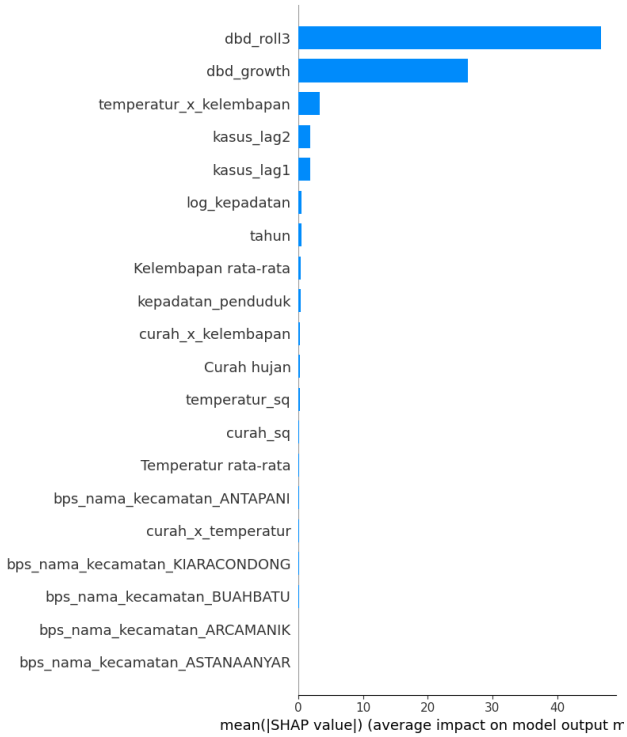

Fig. 10. SHAP bar plot

Figure 10 displays the SHAP feature importance plot based on the absolute mean SHAP value (mean |SHAP value|), which represents the magnitude of each variable's contribution to the overall Random Forest model prediction. The visualization results show that DHF_roll3 is the most dominant feature, followed by DHF_growth, confirming that historical information and the dynamics of DHF case growth play a major role in shaping predictions. The large SHAP values for these two variables indicate that the model is highly dependent on short-term temporal patterns to capture trends in the increase or decrease of cases.

On the other hand, climate and environmental variables, such as the temperature × humidity interaction and case lag variables (kasus_lag1 and kasus_lag2), provide additional contributions with a more moderate level of influence. Meanwhile, demographic variables and spatial indicators based on subdistricts show relatively small importance values, indicating that their role is more contextual than a major driver of predictions. Overall, these results reinforce previous XAI findings that the Random Forest model prioritizes temporal factors as the main determinants, with environmental variables serving as supporting factors in modeling the dynamics of dengue fever spread.

Vol. 1 No.2 January 2026
E-ISSN: 3132-5115
DOI: 10.65780/bima.v1i2.10

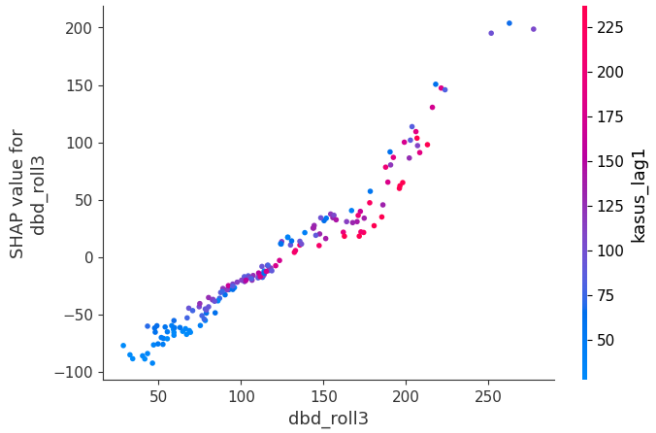

Fig. 11. SHAP dependence plot

Figure 11 is a SHAP dependence plot that shows the relationship between the DHF_roll3 feature value and its contribution to the Random Forest model prediction. The pattern formed shows an almost monotonic relationship, where an increase in the DHF_roll3 value is followed by an increase in the SHAP value, indicating that the DHF case trend in the previous three periods directly contributes to an increase in case predictions in the following period. The coloring of the points based on the kasus_lag1 variable shows an interaction between features, indicating that the influence of DHF_roll3 can be amplified or modulated by the number of cases in the previous period. This finding confirms that the model utilizes a combination of short-term historical information to capture the dynamics of DHF spread in a non-linear and epidemiologically consistent manner.

4.2. Discussions

This research provides a notable contribution by simultaneously addressing predictive performance, learning stability, and model interpretability. Earlier studies have demonstrated the effectiveness of machine learning models, particularly ensemble methods, in capturing non-linear relationships between climate variables and dengue incidence. For example, Tian et al. reported that XGBoost achieved strong predictive performance when applied to meteorological data for dengue prediction, highlighting the advantage of boosting-based algorithms in modeling complex environmental interactions [12]. Similarly, Maulana and Sari showed that ensemble learning approaches, including Random Forest and XGBoost, are capable of producing competitive prediction results for dengue spread modeling [13]. However, these studies primarily focused on performance metrics, with limited discussion on model robustness and interpretability.

Consistent with previous findings, the results of this study confirm that XGBoost achieves superior numerical performance, as indicated by a higher coefficient of determination and lower prediction error compared to Random Forest. This outcome supports prior evidence that boosting-based models are effective in capturing complex and non-linear patterns in epidemiological data [12],[13]. Nevertheless, this research extends beyond conventional performance evaluation by incorporating learning curve analysis. The learning curve results reveal that Random Forest exhibits more stable learning behavior and smaller gaps between training and validation performance, suggesting better generalization capability. This aspect has been largely overlooked in prior dengue prediction studies, which often report high accuracy without sufficient assessment of model stability on unseen data. Another important gap in earlier research relates to the limited application of systematic feature engineering. Many previous studies rely on raw climate or demographic variables, or use them independently, without fully exploiting temporal dependencies and interaction effects [7],[8]. In contrast, this study demonstrates that feature engineering techniques, including temporal lag features, rolling statistics, case growth rates, and climate interaction variables, substantially improve data representation. The prominence of engineered temporal features in the prediction results confirms that these transformations successfully capture the seasonal and temporal dynamics of dengue transmission, which are widely recognized in epidemiological literature [5].

Moreover, while recent studies have begun to incorporate Explainable Artificial Intelligence approaches such as SHAP to improve interpretability, their application in conjunction with ensemble learning models remains limited, particularly at the urban scale [11]. This research addresses this limitation by applying SHAP to the Random Forest model, enabling transparent interpretation of feature contributions. The XAI results indicate that historical case trends are the most influential determinants of dengue predictions, followed by climate interaction variables, findings that are consistent with established knowledge on dengue epidemiology and vector dynamics [3],[5]. By emphasizing interpretability alongside predictive performance, this study responds to concerns raised in the literature regarding the lack of transparency in complex machine learning models used for public health decision-making [9].

Overall, this study advances existing research by demonstrating that high predictive accuracy alone is insufficient for practical implementation in dengue surveillance systems. By integrating feature engineering, comparative ensemble modeling, learning curve analysis, and explainable AI, this research bridges key gaps identified in previous studies [7], [11], [13]. The proposed framework balances accuracy, robustness, and transparency, thereby providing a more reliable and interpretable foundation for the development of data-driven dengue early warning systems at the regional level.

Vol. 1 No.2 January 2026
E-ISSN: 3132-5115
DOI: 10.65780/bima.v1i2.10

5. CONCLUSIONS

This study compares the performance of the XGBoost and Random Forest models in predicting the number of Dengue Hemorrhagic Fever (DHF) cases using the R-square (R^2) and Root Mean Square Error (RMSE) metrics. The evaluation results show that XGBoost produces the best numerical performance with an R^2 value of 0.9520 and an RMSE of 1.6663, while Random Forest obtains an R^2 value of 0.9032 and an RMSE of 2.3748. This difference indicates that XGBoost has a stronger ability to capture complex patterns in the data. However, learning curve analysis shows that Random Forest has a more stable and consistent learning pattern between training and validation data, thus demonstrating better generalization capabilities on unseen data.

In addition to the model's performance achievements, this study also shows that the feature engineering strategy applied successfully improved the quality of data representation. This is reflected in the dominance of temporal features and transformed variables, such as rolling statistics, case growth, and interactions between climate variables, which consistently emerged as major contributors to model predictions based on Explainable Artificial Intelligence (XAI) analysis. The interpretation of the Random Forest model shows that the integration of historical case features and environmental factors produces stable predictions and explanations that are consistent with the epidemiological characteristics of DHF. Thus, this study confirms that the combination of appropriate feature engineering, reliable prediction models, and interpretability approaches provides a strong foundation for the development of accurate and transparent DHF prediction systems to support decision-making at the regional level.

REFERENCES

[1] Pemerintah Provinsi Jawa Barat, "Musim Hujan Tiba, Dinkes Kota Bandung Ajak Masyarakat Tanggap DHF," Diskominfo Kota Bandung.

[2] Redaksi Radar Bandung, "Penurunan DHF 2025 Jadi yang Terendah, Pemkot Bandung Bersiap Hadapi Lonjakan 2026."

[3] E. Abbasi, "*Aedes aegypti* and dengue: insights into transmission dynamics and viral lifecycle," *Epidemiol. Infect.*, vol. 153, p. e88, Aug. 2025, doi: 10.1017/S0950268825100320.

[4] "Dengue and severe dengue," World Healt Organization.

[5] P. Prasad *et al.*, "Influence of climatic factors on the life stages of Aedes mosquitoes and vectorial transmission: A review," *J. Vector Borne Dis.*, vol. 61, no. 2, pp. 158–166, Apr. 2024, doi: 10.4103/jvbd.jvbd_42_24.

[6] A. Kolimenakis *et al.*, "The role of urbanisation in the spread of Aedes mosquitoes and the diseases they transmit—A systematic review," *PLoS Negl. Trop. Dis.*, vol. 15, no. 9, p. e0009631, Sep. 2021, doi: 10.1371/journal.pntd.0009631.

[7] M. Nasir *et al.*, "Machine Learning Approach to Predict the Dengue Cases Based on Climate Factors," *Window of Health : Jurnal Kesehatan*, pp. 203–214, May 2024, doi: 10.33096/woh.vi.1428.

[8] A. Risqa JL, A. Alhaq, and N. Nissa, "Dengue Modeling Using Multiple Regression in Bandar Lampung Province, Indonesia," *Symmetry & Sigma: Journal of Mathematical Structures and Statistical Patterns*, vol. 1, no. 1, pp. 73–86, Jun. 2024, doi: 10.58989/symmerge.v1i1.11.

[9] C. Rudin, C. Chen, Z. Chen, H. Huang, L. Semenova, and C. Zhong, "Interpretable machine learning: Fundamental principles and 10 grand challenges," *Stat. Surv.*, vol. 16, no. none, Jan. 2022, doi: 10.1214/21-SS133.

[10] Z. Liu, Y. Wang, F. Feng, Y. Liu, Z. Li, and Y. Shan, "A DDoS Detection Method Based on Feature Engineering and Machine Learning in Software-Defined Networks," *Sensors*, vol. 23, no. 13, p. 6176, Jul. 2023, doi: 10.3390/s23136176.

[11] Md. S. Rahman and Md. A. B. Shiddik, "Explainable artificial intelligence for predicting dengue outbreaks in Bangladesh using eco-climatic triggers," *Glob. Epidemiol.*, vol. 10, p. 100210, Dec. 2025, doi: 10.1016/j.gloepi.2025.100210.

[12] N. Tian *et al.*, "Precision Prediction for Dengue Fever in Singapore: A Machine Learning Approach Incorporating Meteorological Data," *Trop. Med. Infect. Dis.*, vol. 9, no. 4, p. 72, Mar. 2024, doi: 10.3390/tropicalmed9040072.

[13] H. Maulana and A. Sekar Sari, "Instal : Jurnal Komputer Analysis of Dengue Fever Spread Prediction Using Ensemble Learning Approach with Xgboost and Random", doi: 10.54209/jurnalinstall.v16i03.227.

[14] A. Saleh and R. Mukhtar, "Early Detection of Dengue Hemorrhagic Fever Using Patient Medical Data with Ensemble Learning Methods 1*," *Indonesian Journal of Artificial Intelligence and Data Mining (IJAIDM)*, vol. 8, no. 3, pp. 635–645, 2025, doi: 10.24014/ijaidm.v8i3.38088.

[15] S. Surbakti, Hayatunnufus, and T. Henny Febriana, "Prediction of Dengue Fever in Coastal Areas of North Sumatera (Kuala Namu and Belawan) With Random Forest and Support Vector Machine (SVM) Methods," *Data Science: Journal of Computing and Applied Informatics*, vol. 7, no. 2, pp. 103–110, Jul. 2023, doi: 10.32734/jocai.v7.i2-14355.

[16] P. Probst, M. N. Wright, and A. Boulesteix, "Hyperparameters and tuning strategies for

Vol. 1 No.2 January 2026
E-ISSN: 3132-5115
DOI: 10.65780/bima.v1i2.10

random forest," *WIREs Data Mining and Knowledge Discovery*, vol. 9, no. 3, May 2019, doi: 10.1002/widm.1301.

[17] S.-W. Wang *et al.*, "A case study on the application of a data-driven (XGBoost) approach on the environmental and socio-economic perspectives of agricultural groundwater management," *Agric. Water Manag.*, vol. 318, p. 109729, Sep. 2025, doi: 10.1016/j.agwat.2025.109729.