Optimization of Machine Learning Models for Risk Prediction of DHF Spread to Support Management Strategies in Urban Areas

Yesica Devis[1,*], Muhamadiah[2], Yulanda[3], Yuda Irawan[4], Refni Wahyuni[5]

[1,2]*Department of Public Health, Faculty of Health, Universitas Hang Tuah Pekanbaru, Pekanbaru, Indonesia*

[3,4,5]*Department of Computer Science, Faculty of Computer Science, Universitas Hang Tuah Pekanbaru, Pekanbaru, Indonesia*

(Received: February 18, 2025; Revised: May 06, 2025; Accepted: July 22, 2025; Available online: September 01, 2025)

Abstract

Dengue fever is an endemic disease that poses a serious threat to public health in tropical regions such as Indonesia. Efforts to control this disease require a data-based approach that is able to accurately predict the level of risk so that interventions can be targeted. This study aims to develop a predictive model of DHF risk using ensemble stacking method optimized with Optuna algorithm and integrated into an interactive dashboard based on Streamlit. The dataset used includes environmental, climate, and socio-demographic indicators from 2015 to 2024 with a total of 1,440 data entries. The preprocessing process includes normalization with Standard Scaler, feature selection using LASSO, and label data balancing with the SMOTE method. Model validation was performed using 10-Fold Cross Validation to ensure model generalization to new data. The stacking model is built with three basic algorithms, namely SVM, KNN, and Random Forest, which are combined using Logistic Regression as a meta-learner. The evaluation results show that the model is able to achieve an average accuracy of 97.57%, with high precision, recall, and f1-score values in all three prediction classes (low, medium, high). The ROC-AUC for each class also showed near-perfect performance. The implementation of the model in the Streamlit dashboard allows non-technical users such as health center or health office staff to perform regional risk prediction and obtain data-driven intervention recommendations automatically. This research not only contributes to the development of predictive technology, but also strengthens evidence-based health promotion practices in urban areas. Further research is recommended to integrate IoT-based real-time data and expand the scope of application areas.

Keywords: DHF, Machine Learning, Stacking Ensemble, OPTUNA, Public Health

1. Introduction

Dengue Fever (DHF) is one of the most serious public health challenges globally, especially in tropical and subtropical regions such as Indonesia where the climate is ideal for the development of the Aedes aegypti mosquito vector [1], [2]. Data from the WHO shows that more than 390 million cases of dengue infection occur each year, and about 96 million of them develop into clinical forms that require immediate medical intervention [3], [4]. This situation emphasizes the urgency of a more innovative and data-driven approach in addressing the systemic spread of this disease.

As a DHF-endemic country, Indonesia continues to experience a spike in cases from year to year, including in Riau Province which shows a significant upward trend. In early January 2025, 32 cases were recorded in just the first two weeks, showing an increase compared to the same period the previous year. These dynamics indicate that controlling DHF is not enough with a reactive response, but requires a more adaptive risk prediction system [5], [6]. Environmental factors such as extreme rainfall, high humidity, and temperature changes, coupled with social conditions such as population density and people's living behavior, are variables that interact with each other in increasing the risk of dengue spread [7], [8].

Machine Learning-based predictive approaches are now being applied in an effort to improve the effectiveness of early warning systems against infectious diseases. One prominent method is stacking ensemble, which is an approach of combining several prediction algorithms that produces more accurate models and is resistant to data variations [9], [10]. By integrating basic models such as Decision Tree, K-Nearest Neighbor, and Support Vector Machine into the

***Corresponding author: Yesica Devis (jejes.devis@gmail.com)**

DOI: https://doi.org/10.47738/jads.v6i4.898

This is an open access article under the CC-BY license (https://creativecommons.org/licenses/by/4.0/).
© Authors retain all copyrights

Logistic Regression framework as a meta-model, stacking ensemble is proven to be able to capture the complexity of the relationship between risk factors that affect the spread of DHF [11]. This is an important step in developing evidence-based policies that are right on target, especially in an urban context such as Pekanbaru City.

Recent dengue research has mainly examined environmental, socio-demographic risk factors and community-based interventions [12]. Epidemiological studies have shown that high rainfall, humidity, high temperature, and population density are associated with a surge in dengue cases in tropical urban areas [13], [14]. Other research confirms that the level of education and public awareness of Clean and Healthy Living Behavior plays an important role in the prevention of DHF [15], [16]. Community-based intervention studies through routine health education and vector control have been shown to significantly reduce the incidence of DHF [17], [18], [19], but still limited in terms of long-term effectiveness and less adaptive to changes in epidemiological patterns in various regions [20], [21].

Recent research has begun to utilize Machine Learning methods to support early prediction of dengue cases [22], [23]. Decision Tree and Random Forest based approaches have been widely used [24], [25], [26], but tends to be unstable and less able to cope with epidemiologic data with very complex patterns [26]. The Support Vector Machine (SVM) model shows high prediction performance in certain cases, but is sensitive to the parameters used, making it less flexible to be widely applied [27], [28]. The K-Nearest Neighbor (KNN) method is able to capture the similarity of case patterns between regions [29], but is constrained by the longer computation time when the dataset gets bigger [30], [31].

Recent research indicates that the stacking ensemble method, which integrates several algorithms at once, is able to produce more accurate disease risk predictions than a single model because it combines the advantages of each algorithm [32], [33], [34]. However, the use of stacking ensemble in the context of prediction-based health promotion is still very limited. Based on these weaknesses and limitations, the novelty of this research lies in developing an early prediction model of dengue cases using stacking ensemble by combining machine learning algorithms Decision Tree, KNN, SVM as the base model, and Logistic Regression as a meta-model to improve prediction accuracy so that it can support adaptive, accurate, and data-based public health intervention strategies, especially in Pekanbaru City.

In the field of Public Health, the application of predictive models like this provides a great opportunity to support health promotion and protection strategies that are based on community and real-time data. By integrating prediction results into the dengue management system, health workers can more quickly direct interventions to high-risk areas, conduct targeted education, and optimize resources for vector control in a measured and efficient manner. Therefore, this study aims to develop and test the effectiveness of ensemble stacking models in predicting DHF risk as a basis for strategic decision making in urban areas, especially in Pekanbaru City.

2. Research Methodology

This research focuses on developing a stacking model by integrating various machine learning techniques in order to predict the risk level of DHF based on environmental and social data. The model development process is carried out through a number of systematic stages to ensure high prediction accuracy and relevance to the public health context. The stages of model development can be seen in figure 1.

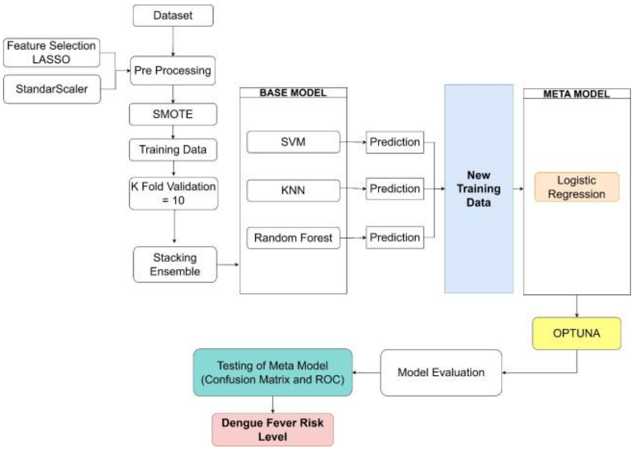


Figure 1. Development of Stacking Model

Figure 1 shows the complete flow of developing a predictive model of DHF risk using the stacked ensemble approach. The process begins with data preprocessing, including standardization and feature selection using LASSO to identify the most significant variables affecting dengue incidence. Next, the SMOTE technique was applied to handle class imbalance, followed by 10-fold cross-validation to ensure model generalization. Three algorithms namely SVM, KNN, and Random Forest used as base learners each generated initial predictions, which were then combined into new training data for the Logistic Regression meta-model. Parameter optimization was performed with the help of Optuna to automatically improve model performance. The results of the final model were evaluated through Confusion Matrix and ROC to measure classification accuracy and performance.

2.1. Dataset

The dataset used in this study is a compilation of several data sources relevant to the context of dengue fever in urban areas. The dataset includes data on dengue cases, weather and climate, environment, and socio-demographics from 2015 to 2024, resulting in a total of 1,440 data entries. The primary data sources include official reports from the Pekanbaru City Health Office for monthly dengue case counts, Meteorology, Climatology, and Geophysical Agency for weather-related indicators such as rainfall, temperature, and humidity, and Environmental Office for environmental data like waste disposal sites and water stagnation reports. In addition, socio-demographic indicators such as unemployment rate and education level were obtained from the Central Bureau of Statistics of Pekanbaru City.

2.2. Pre-processing

The preprocessing stage was conducted to ensure the quality and consistency of the data before it was used in training the prediction model. This process includes checking and handling missing data, normalizing numerical data using StandardScaler, and converting category labels into numerical representations through label encoding. In addition, format and data type alignment were performed to maintain compatibility between variables after integrating datasets from various sources. Next, the feature selection process is carried out using the Least Absolute Shrinkage and Selection Operator (LASSO) method to select the variables that are most relevant to the prediction target. The use of LASSO helps reduce the complexity of the model by eliminating features that contribute less, resulting in a simpler but still accurate model. Optimal preprocessing and feature selection are key in producing a prediction model that is not only accurate, but also able to handle data variations more generally and efficiently. StandardScaler and LASSO feature selection were fitted only on the training set and subsequently applied to the test set. This ensures that no information from the test set influences the model during training.

2.3. Labelling

To support the classification process in developing the prediction model, a DHF risk labeling stage was conducted based on a quantitative approach to the distribution of case variable values. The result of this labeling process divided the data into three risk level categories: low, medium, and high, which were then used as target labels in the training of the Machine Learning model. The labeling results are low = 380, medium = 380, and high = 392. This process is an important foundation in ensuring the model is able to classify accurately and is relevant to the real conditions in the field. The risk labels (low, medium, high) were assigned using a quantile-based approach based on the distribution of historical dengue case counts from 2015 to 2024. This stratification enables proportional comparison across regions and facilitates prioritization in public health interventions, despite the absence of universally standardized thresholds in the literature.

2.4. SMOTE

Synthetic Minority Oversampling Technique (SMOTE) was used in this study to overcome the problem of data imbalance between classes in DHF risk labeling. Before SMOTE, the data distribution was very unbalanced, where the High class had much more data than other classes such as Low or Medium. This imbalance could potentially cause bias in the training of machine learning models, so that the models tend to ignore minority classes. By applying SMOTE, synthetic samples of the minority class are generated based on the closest features, so that the amount of data for each class is balanced. This process helps the model to learn fairly across classes, improving prediction accuracy and generalization ability. The table 1 shows the comparison of the amount of data between classes before and after SMOTE.

Table 1. Comparative Analysis of the Dataset Before and After Applying SMOTE

Risk Level	Before SMOTE	After SMOTE
Low	380	392
Medium	380	392
High	392	392

Table 1 shows the comparison of data distribution based on DHF risk level before and after the SMOTE process. It can be seen that before SMOTE, the High class has the most dominant amount of data with 392 samples, while the Low and Medium classes have only 380 samples each. This imbalance in the amount of data can affect the performance of the model in recognizing patterns from the minority class, so it tends to produce biased predictions towards the majority class. After applying SMOTE, all classes are balanced to 392 samples, so the amount of data in each class becomes proportional. With this balanced data distribution, the model training process can be fairer for all classes, improving the ability of the model to detect DHF risk evenly and accurately. To prevent information leakage, SMOTE was applied only within the training folds during each iteration of cross-validation. This ensures that synthetic data generated by SMOTE did not influence the validation set, thereby preserving the integrity and generalizability of the model evaluation.

2.5. K-Vold Cross Validation

K-Fold Cross Validation is a technique used to more accurately evaluate the performance of machine learning models by dividing the dataset into subsets or folds of equal size [35], [36]. In this study, a 10-Fold Cross Validation scheme (K=10) is used, where the dataset of 1,440 data is divided into 10 subsets. At each iteration, one subset is used as test data and the remaining nine subsets are used as training data. This process is repeated 10 times so that each subset acts as test data once. The selection of the K=10 value is based on the balance between bias and variance, computational efficiency, and support from empirical evidence. The value of K=10 is considered optimal because it is able to reduce high variance at small K values and minimize bias at large K values. This technique provides efficient model validation without excessive computational burden, especially for medium-sized datasets such as in this study. By using 10-Fold Cross Validation, the model is trained and tested on all parts of the data, making the performance evaluation more stable and able to provide good generalization to new data that has never been seen before.

2.6. Stacking Ensemble Model

Stacking is one of the ensemble techniques in machine learning that combines predictions from multiple base models using meta models to produce more accurate final predictions [37]. The stacking ensemble method used in this study integrates multiple base learners to improve classification performance through a two-layer architecture. Let the training dataset be defined as $D = \{(x_i, y_i)\}_{i=1}^{n}, where\, x_i \in R^d$ is a feature vector and y_i is the $\{h_1(.), h_2(.), \dots, h_M(.)\}$, such as Random Forest, SVM, and KNN, each producing individual predictions $h_m(x_i)$. These outputs are concatenated into a meta-feature vector $z_i = [h_1(x_i), h_2(x_i), \dots., h_M(x_i)]^T$, which serves as the input to the meta-learner $h_{meta}(.)$. The meta-learner, modeled using Logistic Regression, is further optimized using the Optuna framework. Optuna automatically explores the hyperparameter space Θ to maximize cross-validated accuracy through a Bayesian optimization process, defined as $\theta^* = \arg max_{\theta \in \Theta}$ Accuracy$(h_{meta}(z_i; \theta))$. Furthermore, a passthrough mechanism is applied, allowing the original input features x_i to be included alongside the base learner predictions, forming an extended feature vector $z_i' = [x_i; z_i]$. The final prediction is thus computed as $\hat{y}_i = h_{meta}(z_i')$.

This formulation allows the stacking ensemble to leverage the complementary strengths of heterogeneous models while ensuring the meta-learner is optimally tuned for the specific classification task, in this case, predicting dengue risk levels. The architecture has proven to enhance predictive performance over individual models, as demonstrated by the evaluation results. Table 2 in the following literature review section summarizes a number of previous studies that have used stacking techniques in the development of machine learning models.

Table 2. The Previous Research Related to Stacking

Researcher	Based Model	Meta Model	Accuracy
Hasan [38]	SVM, XGBoost, Artificial Neural Networks (ANN)	Logistic Regression	95.00%
Hou [39]	SVM, KNN, Random Forest (RF), Gradient Boosting Decision tree (GBDT)	GBDT	93.10%
Liu [40]	Logistic Regression (LR), RF, Extreme Gradient Boosting (XGBoost)	Logistic Regression	95.00%
Zheng [41]	LR, SVM, KNN, Decision tree (DT), RF, XGBoost, AdaBoost	RF	93.80%
Kshatri [42]	SVM, J48, Naïve Bayes, Bagging, Random Forest	SVM	94.50%

Based on table 2, it can be seen that stacking techniques have been widely applied in various studies to improve the prediction accuracy of machine learning models. Various combinations of base models such as SVM, KNN, Random Forest, Logistic Regression, XGBoost, and Naive Bayes are used with various meta models, such as Logistic Regression, Gradient Boosting, Random Forest, and SVM. The accuracy results obtained are quite high, ranging from 93.10% to 95.00%. This shows that the utilization of ensemble stacking models is generally able to improve model performance compared to the use of a single model. These findings support the use of stacking as an efficient approach in managing data complexity and improving model generalization, as well as providing flexibility in the selection of suitable algorithms for public health data-based prediction problems. However, all studies summarized in table 2 have not implemented data balancing techniques such as SMOTE and have not integrated hyperparameter optimization using Optuna. The model proposed in this study aims to overcome these limitations and significantly improve prediction accuracy as shown in table 3.

Table 3. Proposed Stacking Model

Base Model	Meta Model	Optimization
SVM, KNN, Random Forest	Logistic Regression	Standard Scaler, LASSO, SMOTE, Hyperparameter Tuning (optuna)

Table 3 shows the design of the stacking model proposed in this study, which integrates a combination of SVM, KNN, and Random Forest base models combined through a Logistic Regression meta model. Unlike previous studies that only rely on the basic stacking structure, this model strengthens the prediction process through a series of systematic optimization stages, namely normalization using Standard Scaler, LASSO-based feature selection, class balancing using SMOTE, and automatic hyperparameter tuning with the help of Optuna. These four stages are designed to overcome classic challenges in epidemiological data such as skewed class distribution, noise, and overfitting, thus significantly improving the generalizability of the model. With this approach, the resulting model is expected to be more stable, accurate, and relevant in supporting data-driven public health intervention strategies in urban areas.

2.7. Model Evaluation

The model evaluation in this study was conducted comprehensively using evaluation metrics including accuracy, confusion matrix, precision, recall, F1-score, and ROC-AUC. These metrics were chosen because they are able to provide a comprehensive overview of the model's performance in accurately classifying DHF risk levels. Confusion matrix is used to see the distribution of correct and incorrect classifications in each class [29], whereas precision and recall emphasize the model's ability to avoid misclassification, especially for high-risk categories that are of great public health importance. F1-score is an indicator of the balance between precision and recall [32]. In addition, ROC curves and AUC values are used to measure the model's ability to distinguish between risk classes as a whole [33].

3. Results and Discussion

The dataset used in this study consists of a combination of relevant data sources that reflect the environmental, social and weather factors that influence the incidence of DHF. Overall, the dataset includes 1,440 entries collected from

2015 to 2024 and includes data per sub-district in urban areas. Data components include dengue case records, rainfall data, average temperature, humidity, as well as socio-demographic information such as population density and education levels. This data is designed to represent the complex actual conditions in the field, while providing a solid foundation for the development of an adaptive and accurate DHF risk prediction model. The integration of various types of data over a long period of time is expected to increase the generalization capacity of the model and provide predictive results that are contextual and relevant to local public health dynamics. After applying LASSO to the normalized data, table 4 show the results were obtained.

Table 4. LASSO Coefficients of Features

No	Features	LASSO Coefficient
1	rainfall	-13.81
2	dengue_cases	-10.3
3	water_puddles	-4.67
4	illegal_dump_sites	-2.2
5	humidity	-1.75
6	education_level	-0.72
7	unemployment_rate	-0.6
8	average_temperature	-0.47
9	fogging_frequency	0.74

Table 4 presents the results of feature selection using the LASSO method after normalizing the data. It can be seen that rainfall and dengue_case_count have the largest negative coefficients, -13.81 and -10.3 respectively, indicating a dominant contribution to dengue risk prediction. Meanwhile, fogging_activity is the only variable with a positive coefficient (0.74), indicating a unidirectional correlation to risk. Other variables such as standing_water, waste_dump, and population_density also showed relevant contributions although smaller. These results confirm the importance of environmental and climatic factors in influencing the surge in dengue cases, in line with previous epidemiological studies on the spread of vector-based diseases.

This feature selection not only considers statistical significance, but is also based on the principles of disease ecology and the social determinants of health approach. The retained features, such as education_level and unemployment_rate, reflect the community's capacity to take promotive and preventive actions. Thus, the resulting predictive model has methodological advantages as well as practical benefits in planning data-driven health interventions. This strengthens the role of machine learning in supporting evidence-based public health strategies, especially for dengue control in urban areas. Furthermore, socio-demographic variables such as education level and unemployment rate were found to influence the predictive model. In public health contexts, lower education is often associated with limited awareness of dengue prevention practices, while higher unemployment may indicate weaker household sanitation and access to resources, both of which contribute to increased vulnerability to DHF outbreaks. These variables serve as indirect indicators of community resilience and behavioral risk. Based on the results of the labeling process on the dataset, the distribution of DHF risk classes is obtained as presented in the figure 2.

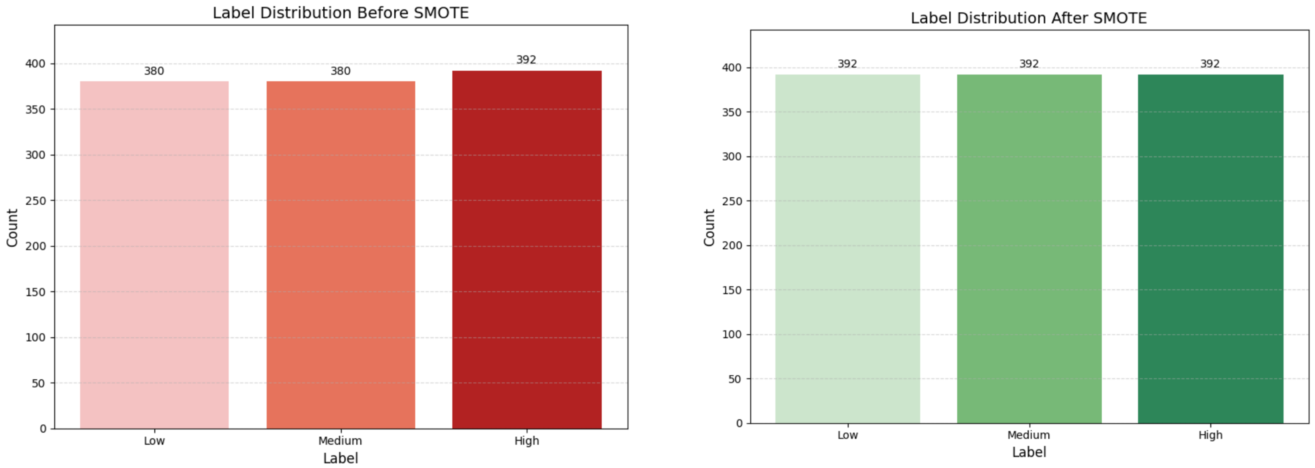


Figure 2. Label Distribution Chart Before and After SMOTE

The graph shows a comparison of the distribution of DHF risk level labels before and after the application of the SMOTE method. Before balancing, it can be seen that the High label has a slightly larger amount of data 392 compared to the Low and Medium labels which only amounted to 380 each, indicating a minor imbalance between classes. After the SMOTE process was applied, all label categories were equalized to 392, resulting in a balanced data distribution. These results show that the SMOTE technique is effective in overcoming the data imbalance problem, which is important for improving the performance of the classification model in detecting all risk categories fairly and accurately.

After the data balancing process using SMOTE, the classification stage is performed by applying the stacking ensemble technique. This model integrates several basic algorithms such as SVM, KNN, and Random Forest, with Logistic Regression as a meta-model that is optimized through Optuna to improve prediction performance. The logistic regression model was optimized with two key hyperparameters: the regularization strength (C), searched within a log-uniform distribution ranging from 1e-3 to 10, and the solver type, selected from liblinear and lbfgs. A total of 30 trials were performed, each evaluated using 3-fold cross-validation. The best-performing combination was selected to build the final stacking ensemble model. Table 5 show the classification report results of the stacking model with the Logistic Regression meta-model that has been optimized using Optuna on data that has been balanced with SMOTE.

Table 5. Classification Report

Class	Precision	Recall	F1-Score	Support
Low	1.00	0.98	0.99	95
Medium	0.96	0.99	0.97	95
High	0.99	0.98	0.98	98
Accuracy			0.98	288
Macro Avg	0.98	0.98	0.98	288
Weighted Avg	0.98	0.98	0.98	288

The classification table above shows that the developed stacking model, with the Logistic Regression meta model optimized using Optuna, achieved very high performance overall. The Low class obtained the perfect precision (1.00) and the highest F1-score (0.99), indicating the model was able to identify this class without significant error. Although the Medium class had a slightly lower precision (0.96), its recall reached 0.99, indicating the model did not miss many predictions for this class. Consistent macro average and weighted average values of 0.98 indicate a balanced performance of the model across all classes, making it valid for application in the context of DHF risk prediction. This confirms that the use of Optuna for hyperparameter tuning significantly contributes to improving the accuracy and stability of model predictions. Figure 3 displays the confusion matrix results illustrating the classification capability of the stacking model with a very minimal misclassification rate.

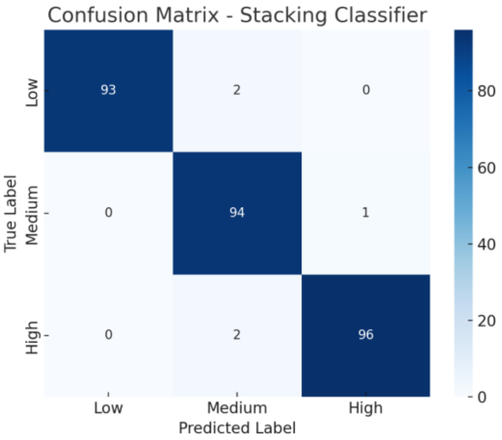


Figure 3. Confusion Matrix

Based on the confusion matrix of the stacking model test results, it is known that the model has very good classification performance for all three classes, namely Low, Medium, and High. The model is able to classify 93 out of 95 Low class data correctly, only wrong in 2 instances. For the Medium class, the accuracy is higher with 94 out of 95 data successfully predicted correctly, and only 1 data misclassified. Meanwhile, the High class shows the highest accuracy, with 96 out of 98 data successfully classified correctly. This very low prediction error rate indicates that the stacking model equipped with hyperparameter optimization is able to distinguish dengue risk categories very effectively, and demonstrates stability and accuracy in multiclass classification. The confusion matrix confirms the model's high performance, as it produces minimal misclassifications across classes. This ensures that the model can serve as a dependable decision-support tool in the field.

These results indicate that the proposed model not only achieves high accuracy but also maintains strong precision and recall across all risk categories, making it reliable for supporting public health surveillance and early dengue prevention strategies in urban communities. The evaluation metrics reported are based on cross-validation that integrates SMOTE within each fold, ensuring realistic estimation of model performance. In addition to accuracy, the evaluation also includes Cohen's Kappa and Matthews Correlation Coefficient (MCC) to assess the model's performance under class imbalance. The obtained Kappa score of 0.93 and MCC score of 0.91 suggest strong agreement and robust classification performance across the three risk categories. Next, figure 4 is the result of the ROC:

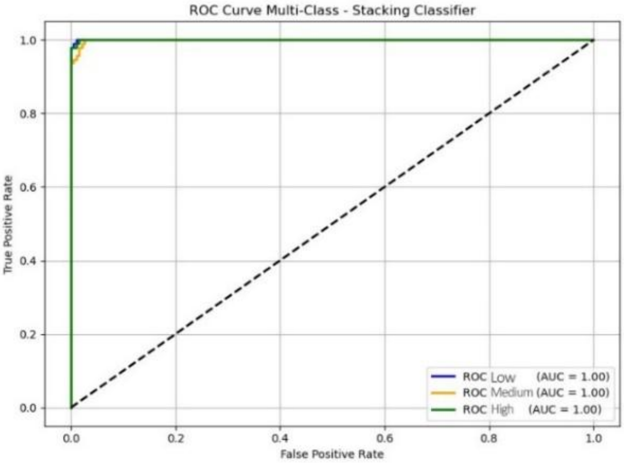


Figure 4. ROC Graph

The Multi-Class ROC Curve image on the stacking classifier model shows very optimal performance in distinguishing the three classes of DHF risk levels, namely Low, Medium, and High. This can be seen from the Area Under Curve (AUC) value which reaches 1.00 for the three classes, which indicates that the model's discriminatory ability is at a perfect level. The ROC curve of each class is very close to the upper left corner of the graph, indicating a high True

Positive Rate and a very low False Positive Rate. With these results, it can be concluded that the developed stacking model has very good classification performance in identifying each category of DHF risk, and is able to provide accurate and reliable predictions for data-based public health interventions. The ROC curve shows an AUC close to 1.00 for each class, demonstrating that the model is capable of distinguishing between low, medium, and high dengue risk levels with excellent discrimination ability. Although the ROC AUC scores for all classes are reported as 1.00, these results are not due to overfitting. The model performance was evaluated through 10-fold cross-validation, and the use of LASSO feature selection, standard scaling, and SMOTE ensured that overfitting and data leakage risks were minimized. K-Fold Cross Validation is an evaluation method that aims to measure the consistency and generalization ability of the model to different data. Table 6 show results of the K-Fold Cross Validation for the stacking model can be seen in table 6 below:

Table 6. 10-Fold Cross Validation Results

Fold	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Fold 6	Fold 7	Fold 8	Fold 9	Fold 10	**Average**
Accuracy	0.9792	1.00	0.9861	0.9653	0.9861	0.9653	0.9861	0.9583	0.9514	0.9792	0.9757

Based on the results of the 10-Fold Cross Validation evaluation shown in table 6, the stacking model shows consistent performance with high accuracy values in each fold, which are in the range of 0.9514 to 1.0000. The average accuracy of 0.9757 indicates that the model has excellent generalization capabilities for data that has never been seen before. Despite achieving a high accuracy of 97.57%, several precautions were taken to prevent overfitting, including 10-fold cross-validation, SMOTE for handling class imbalance, and hyperparameter tuning using Optuna. These techniques ensure that the performance reflects the model's generalizability rather than memorization of training data. Figure 5 shows a comparison of the average accuracy (CV Mean) of various classification models, including individual models (Random Forest, SVM, KNN), stacking ensemble, and stacking optimized using Optuna.

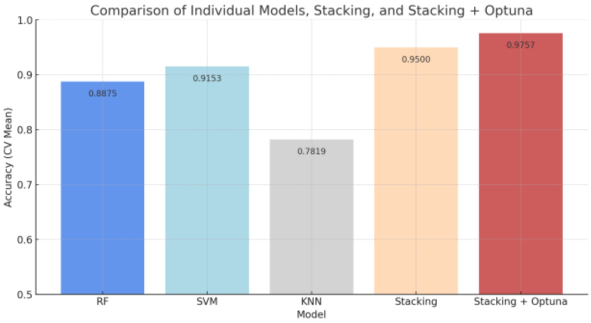


Figure 5. The Result of Stacking Comparison

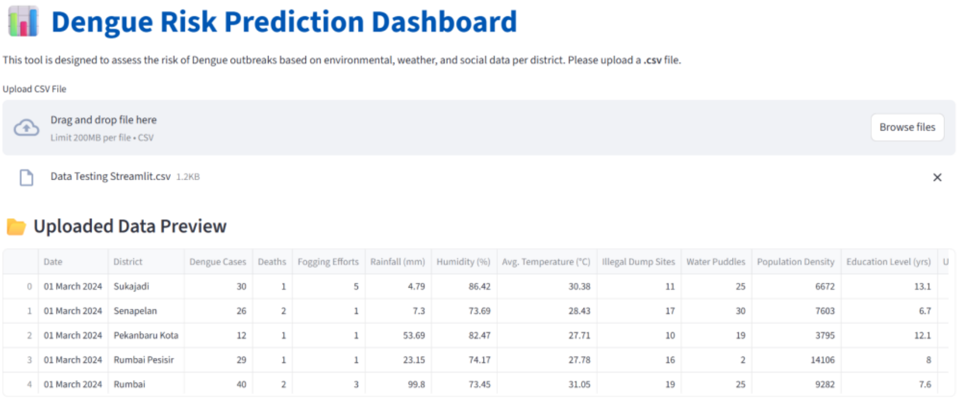
The results show that the stacking model has a higher performance (0.9500) than the individual model, with KNN recording the lowest accuracy (0.7819). After hyperparameter tuning using Optuna, the model accuracy increased to 0.9757, making it the model with the best performance overall. While the proposed stacking ensemble model demonstrates notable gains in prediction accuracy, it inherently carries higher model complexity due to the integration of multiple base learners and a meta-model. Inference time is also relatively longer compared to single classifiers like Decision Tree or Naïve Bayes. However, this trade-off is considered acceptable in public health applications where predictive robustness and multiclass handling are prioritized over millisecond-level prediction latency. This study was then compared with previous studies and showed superior results. Table 7 presents a comparison with previous studies.

Table 7. The Comparison with Previous Research

Researcher	Based Model	Meta Model	Accuracy
Hasan [38]	SVM, XGBoost, ANN	Logistic Regression	95.00%
Hou [39]	SVM, KNN, RF, GBDT	GBDT	93.10%
Liu [40]	LR, RF, XGBoost	Logistic Regression	95.00%
Zheng [41]	LR, SVM, KNN, DT, RF, XGBoost, AdaBoost	RF	93.80%
Kshatri [42]	SVM, J48, Naïve Bayes, Bagging, Random Forest	SVM	94.50%
This Research			97.57%

As shown in table 7, the proposed model achieved the highest accuracy of 97.57%, outperforming previous studies using stacking ensemble methods, which reported accuracies between 93.10% and 95.00%. This improvement was driven by the integration of stacking ensemble, LASSO-based feature selection, SMOTE, and hyperparameter tuning with Optuna. Compared to prior studies that predominantly relied on standalone models without advanced optimization, our research demonstrates the advantage of combining these techniques to significantly enhance predictive performance and support the development of an adaptive, data-driven infectious disease risk prediction system.

The DHF risk prediction model integrated into the Streamlit application was developed to classify the risk level per sub-district into three categories: Low, Medium, and High while providing contextual public health intervention recommendations. This dashboard (see figure 6) utilizes environmental, social, and weather data uploaded in .csv format, and is designed with a simple and intuitive interface, making it easy to operate by non-technical users such as Puskesmas or Health Office officers. The app's automated recommendation feature guides public health principle-based strategies such as early detection, data-driven education, community empowerment, and strengthening local surveillance.



📊 Dengue Risk Prediction Dashboard

This tool is designed to assess the risk of Dengue outbreaks based on environmental, weather, and social data per district. Please upload a **.csv** file.

Upload CSV File

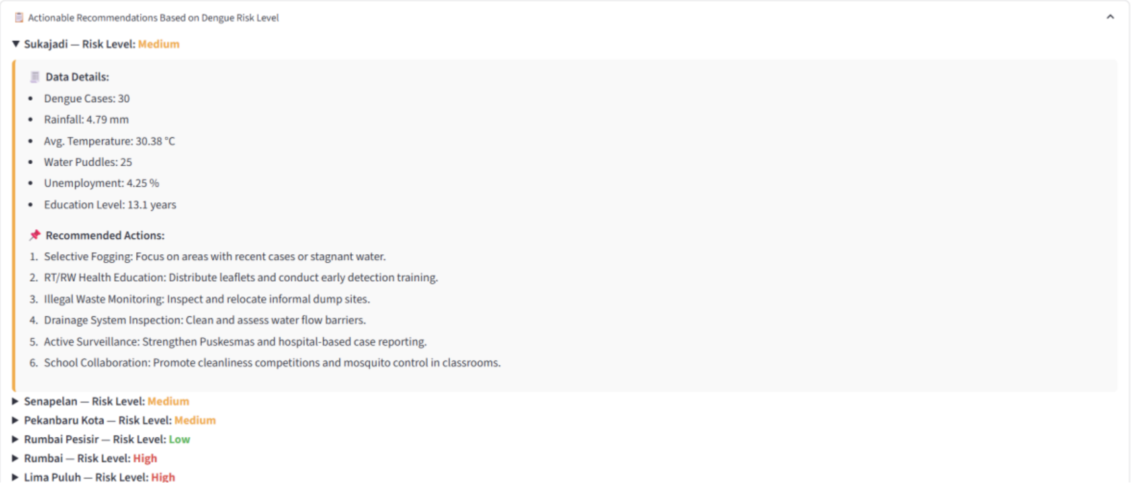
Data Testing Streamlit.csv 1.2KB

📁 Uploaded Data Preview

	Date	District	Dengue Cases	Deaths	Fogging Efforts	Rainfall (mm)	Humidity (%)	Avg. Temperature (°C)	Illegal Dump Sites	Water Puddles	Population Density	Education Level (yrs)	U
0	01 March 2024	Sukajadi	30	1	5	4.79	86.42	30.38	11	25	6672	13.1	
1	01 March 2024	Senapelan	26	2	1	7.3	73.69	28.43	17	30	7603	6.7	
2	01 March 2024	Pekanbaru Kota	12	1	1	53.69	82.47	27.71	10	19	3795	12.1	
3	01 March 2024	Rumbai Pesisir	29	1	1	23.15	74.17	27.78	16	2	14106	8	
4	01 March 2024	Rumbai	40	2	3	99.8	73.45	31.05	19	25	9282	7.6	

Figure 6. Initial Interface Display of Dengue Fever Risk Prediction Application

The image shows the initial display of the interface of the Machine Learning-based DHF risk prediction application built using Streamlit. Users can upload CSV files containing environmental, weather, and social indicator data, which are then automatically displayed in an interactive table for further analysis. From the uploaded file, predictions are made using a machine learning model and the results are in the form of predictions and recommendations in figure 7 below:



Actionable Recommendations Based on Dengue Risk Level

▼ Sukajadi — Risk Level: Medium

Data Details:

- Dengue Cases: 30
- Rainfall: 4.79 mm
- Avg. Temperature: 30.38 °C
- Water Puddles: 25
- Unemployment: 4.25 %
- Education Level: 13.1 years

📌 Recommended Actions:

1. Selective Fogging: Focus on areas with recent cases or stagnant water.
2. RT/RW Health Education: Distribute leaflets and conduct early detection training.
3. Illegal Waste Monitoring: Inspect and relocate informal dump sites.
4. Drainage System Inspection: Clean and assess water flow barriers.
5. Active Surveillance: Strengthen Puskesmas and hospital-based case reporting.
6. School Collaboration: Promote cleanliness competitions and mosquito control in classrooms.

▶ Senapelan — Risk Level: Medium
▶ Pekanbaru Kota — Risk Level: Medium
▶ Rumbai Pesisir — Risk Level: Low
▶ Rumbai — Risk Level: High
▶ Lima Puluh — Risk Level: High

Figure 7. Display of Recommended Actions Based on Predicted Dengue Fever Risk Levels per Sub-district

Figure 7 shows the display of the recommended action feature based on the results of the DHF risk level prediction per sub-district in the application. Each sub-district is grouped into three risk categories (low, medium, high) which are marked with different colors. For example, for areas with medium risk such as Sukajadi, the system provides recommendations such as selective fogging, RT/RW education, and monitoring of illegal TPS and water channels, which are arranged to facilitate follow-up by health workers or local policy makers.To support public health implementation, the application was developed with basic privacy and security safeguards, including HTTPS support, non-identifiable data processing, and adherence to standard public health data protection principles.

Although the dataset used in this study spans a 10-year period, the data was pooled and treated as a single cross-sectional set. As a result, the model does not capture temporal fluctuations in dengue incidence. Future studies are encouraged to implement temporal validation strategies such as time-series cross-validation or rolling forecasting origin to account for seasonal patterns and year-to-year variability in epidemiological trends. One important limitation of this study is the relatively small size of the dataset 1,440 records, which may limit the model's capacity to perform full generalization in a multiclass setting. However, this limitation is addressed through robust preprocessing, feature selection via LASSO, data balancing using SMOTE, and the application of ensemble methods to optimize performance within the constraints of the available public health data. Future extensions may include SHAP or decision boundary plots to enhance explainability, especially in public health decision-making contexts.

4. Conclusion

This study successfully developed a predictive model of DHF risk based on the stacking ensemble method integrated in an interactive dashboard based on Streamlit. The model development process begins with data preprocessing using Standard Scaler for normalization, and feature selection using the LASSO method to filter the most relevant variables. To overcome class imbalance in the data, the SMOTE technique is used so that the label distribution is balanced and the model is not biased towards the majority class. Model validation was carried out using 10-Fold Cross Validation to ensure that the evaluation results have good generalization. The final model combines three base models (SVM, KNN, and Random Forest) and one Logistic Regression meta model optimized using Optuna, resulting in an average accuracy of 97.57%. This model is not only superior in predictive performance, but is also able to provide contextual risk-based intervention recommendations, thus supporting more targeted public health decision-making. The prediction results can be used by the Health Office and Health Centers to plan effective health promotion and vector control strategies. This predictive framework can assist public health decision-makers in prioritizing interventions, allocating resources, and designing adaptive vector control programs based on real-time data and spatial risk mapping. Further research can be directed towards developing real-time prediction systems with integration of IoT data and sensors for the collection of meteorological data such as rainfall, and air humidity, as well as the application of the model in different geographical areas to test the robustness and adaptability of the proposed approach. This study is limited by the absence of external validation using datasets from other urban areas outside Pekanbaru. Although the model shows strong internal performance, further validation is required to assess its generalizability in different epidemiological and environmental contexts across Indonesia.

5. Declarations

5.1. Author Contributions

Conceptualization: Y.D., M., Y., Y.I., R.W.; Methodology: Y.D., M., Y.; Software: M., Y.I.; Validation: Y., R.W.; Formal Analysis: Y.D.; Investigation: M., Y.D.; Resources: R.W., Y.I.; Data Curation: Y.D., Y.; Writing – Original Draft Preparation: Y.D.; Writing – Review and Editing: M., Y., Y.I., R.W.; Visualization: Y.D.; All authors have read and agreed to the published version of the manuscript.

5.2. Data Availability Statement

The data presented in this study are available on request from the corresponding author.

5.3. Funding

This research is funded by the Directorate of Research and Community Service, Directorate General of Research and Development, Ministry of Higher Education, Science, and Technology with the Regular Fundamental Research Scheme in 2025, in accordance with Research Contract Number: 010/LL17/DT.05.00/PL/2025.

5.4. Institutional Review Board Statement

Not applicable.

5.5. Informed Consent Statement

Not applicable.

5.6. Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] S. Masyeni, I. M. W. Wardhana, and F. Nainu, "Cytokine profiles in dengue fever and dengue hemorrhagic fever: A study from Indonesia.," *Narra J*, vol. 4, no. 1, pp. 1-11, Apr. 2024, doi: 10.52225/narra.v4i1.309.

[2] Bur, R., Suwarto, S., Pohan, H.T., "Early intervention of 5% albumin shown superior control of vascular integrity and function compared to ringer's lactatein hospitalized adult with grade I & II Dengue hemorrhagic fever: a multicenter randomized controlled trial in Indonesia," *Trop. Dis. Travel Med. Vaccines*, vol. 10, no. 1, pp. 1-12, 2024, doi: 10.1186/s40794-024-00230-3.

[3] M. P. Patel, V. M. Oza, H. B. Tanna, A. D. Khadela, P. D. Bharadia, and J. K. Patel, "Current Perspectives in Dengue Hemorrhagic Fever," in *Rising Contagious Diseases*, vol 5, no. 3, pp. 72–86, 2024.

[4] Riaz, M., Harun, S.N.B., Mallhi, T.H.., "Evaluation of clinical and laboratory characteristics of dengue viral infection and risk factors of dengue hemorrhagic fever: a multi-center retrospective analysis," *BMC Infect. Dis.*, vol. 24, no. 1, pp. 1-13, 2024, doi: 10.1186/s12879-024-09384-z.

[5] S. J. Hasani, Sgroi. Giovanni., "Recent advances in the control of dengue fever using herbal and synthetic drugs," *Heliyon*, vol. 11, no. 3, pp. 1.13, 2025, doi: 10.1016/j.heliyon.2025.e41939.

[6] A. Bernardin, T. Masrour, B. Partridge, A. J. M. Martin, A. Kelly, and T. Perez-Acle, "Proper pandemic preparedness requires an integrated cross-regional effort, the case of the ECLIPSE consortium in America: a narrative review," *J. Heal. Popul. Nutr.*, vol. 44, no. 1, pp. 1.12, 2025, doi: 10.1186/s41043-025-00850-1.

[7] J. Islam, F. D. Frentiu, G. J. Devine, H. Bambrick, and W. Hu, "A State-of-the-Science Review of Long-Term Predictions of Climate Change Impacts on Dengue Transmission Risk," *Environ. Health Perspect.*, vol. 133, no. 5, pp. 1-13, 2025, doi: 10.1289/EHP14463.

[8] F. Feng, Y. Ma, Y. Zhao, Z. Liu, R. Zhang, and Z. Wan, "Assessment of Global Dengue Transmission Risk Under Future Climate Scenarios," *Earth's Futur.*, vol. 13, no. 7, pp. 1.11, 2025,

[9] P. A. D. Amiri and S. Pierre, "An Ensemble-Based Machine Learning Model for Forecasting Network Traffic in VANET," *IEEE Access*, vol. 11, no. March, pp. 22855–22870, 2023, doi: 10.1109/ACCESS.2023.3253625.

[10] I. D. Mienye and Y. Sun, "A Survey of Ensemble Learning: Concepts, Algorithms, Applications, and Prospects," *IEEE Access*, vol. 10, no. September, pp. 99129–99149, 2022, doi: 10.1109/ACCESS.2022.3207287.

[11] Reshan, M.S.A.; Amin, S.; Zeb, M.A.; Sulaiman, A.; Alshahrani, H.; Azar, A.T.; Shaikh, A., "Enhancing Breast Cancer Detection and Classification Using Advanced Multi-Model Features and Ensemble Machine Learning Techniques," *Life*, vol. 13, no. 10, pp. 1-12, 2023, doi: 10.3390/life13102093.

[12] S. Ghimire and S. Pangeni, "A mixed method evaluation of knowledge, attitude and practice on dengue fever among Lalitpur Metropolitan City residents: a cross-sectional investigation," *BMC Infect. Dis.*, vol. 24, no. 1, pp. 1-11, 2024, doi: 10.1186/s12879-024-10025-8.

[13] N. A. M. H. Abdullah, N. C. Dom, S. A. Salleh, H. Salim, and N. Precha, "The association between dengue case and climate: A systematic review and meta-analysis.," *One Heal. (Amsterdam, Netherlands)*, vol. 15, no. 2, pp. 1-11, Dec. 2022, doi: 10.1016/j.onehlt.2022.100452.

[14] P. S. Singh and H. K. Chaturvedi, "Socio-ecological predictors of dengue in high incidence area of Delhi, India," *Sci. Rep.*, vol. 14, no. 1, pp. 1-11, 2024, doi: 10.1038/s41598-024-67909-7.

[15] M. Hamed, "Knowledge, attitude, and practices toward dengue fever among the public: a cross-sectional study in the Western region of Saudi Arabia," *Front. Public Heal.*, vol. 12, no.2, pp. 1-10, 2024, doi: 10.3389/fpubh.2024.1327427.

[16] M. N. Chaudhary., “Assessing the basic knowledge and awareness of dengue fever prevention among migrant workers in Klang Valley, Malaysia,” *PLoS One*, vol. 19, no. 2, pp. 1-12, Feb. 2024,

[17] S. Fahri and S. Suhermanto, “Enhancing Student Knowledge in Dengue Hemorrhagic Fever Control Through Educational Modeling,” *Heal. Educ. Heal. Promot.*, vol. 12, no. 1, pp. 1-10, 2024, doi: 10.58209/hehp.12.1.37.

[18] F. Aslam., “Effect of Educational Interventions on Awareness of Dengue Fever and Its Preventive Measures among High School Students: Educational Interventions On Awareness of Dengue Fever,” *NURSEARCHER (Journal Nurs. Midwifery Sci.*, vol. 4, no. 03 SE-Original Articles, pp. 24–29, Sep. 2024, doi: 10.54393/nrs.v4i03.114.

[19] Dapari, R., Jumidey, A.Q., Manaf, R.A., “School-based health education effect on knowledge, attitude, and practices of dengue prevention among school children: a systematic review,” *Discov. Soc. Sci. Heal.*, vol. 5, no. 1, p0. 1-10, 2025, doi: 10.1007/s44155-025-00181-w.

[20] López-Saleme Rossana, Escobar-Velásquez Katty, and Barajas-Lizarazo Mayra, “Knowledge, Attitudes, and Practices for the Prevention and Vector Control of Dengue in a Colombian Rural Population: A Mixed Method Study,” *SAGE Open Nurs.*, vol. 11, no Jan, p.1-12, 2025, doi: 10.1177/23779608241302713.

[21] Jamal, M.K., Sanaei, B., Naderi, M., “Investigating the recent outbreak of dengue fever in Iran: a systematic review,” *Egypt. J. Intern. Med.*, vol. 37, no. 1, pp. 1-12, 2025, doi: 10.1186/s43162-025-00411-2.

[22] Tian, N.; Zheng, J.-X.; Li, L.-H.; Xue, J.-B.; Xia, S.; Lv, S.; Zhou, X.-N., “Precision Prediction for Dengue Fever in Singapore: A Machine Learning Approach Incorporating Meteorological Data,” *Trop. Med. Infect. Dis.*, vol. 9, no. 4, 2024, doi: 10.3390/tropicalmed9040072.

[23] Madewell, Z.J., Rodriguez, D.M., Thayer, M.B., “Machine learning for predicting severe dengue in Puerto Rico,” *Infect. Dis. Poverty*, vol. 14, no. 1, pp. 1-12, 2025, doi: 10.1186/s40249-025-01273-0.

[24] I. P. Hawa, S. S. Prasetiyowati, and Y. Sibaroni, “Classification Prediction of Dengue Fever Spread Using Decision Tree with Time-Based Feature Expansion,” *Int. J. Inf. Commun. Technol.*, vol. 10, no. 2, pp. 225–241, 2025, doi: 10.21108/ijoict.v10i2.1026.

[25] Anam, M.K., Van FC, L.L., Hamdani, H., Rahmaddeni, R., Junadhi, J., Firdaus, M.B., Syahputra, I. and Irawan, Y., “Sara Detection on Social Media Using Deep Learning Algorithm Development ,” *J. Appl. Eng. Technol. Sci.*, vol. 6, no. 1 SE-Articles, pp. 225–237, Dec. 2024, doi: 10.37385/jaets.v6i1.5390.

[26] D. A. Tuan and T. N. Dang, “Leveraging Climate Data for Dengue Forecasting in Ba Ria Vung Tau Province, Vietnam: An Advanced Machine Learning Approach.,” *Trop. Med. Infect. Dis.*, vol. 9, no. 10, pp. 1-10, Oct. 2024, doi: 10.3390/tropicalmed9100250.

[27] A. S. Khan., “Integrating BERT Embeddings with SVM for Prostate Cancer Prediction,” in *2024 6th International Conference on Electrical Engineering and Information & Communication Technology (ICEEICT)*, vol. 10534547, no. May, pp. 1–6, 2024. doi: 10.1109/ICEEICT62016.2024.10534547.

[28] S. Samantaray and A. Sahoo, “Prediction of flow discharge in Mahanadi River Basin, India, based on novel hybrid SVM approaches,” *Environ. Dev. Sustain.*, vol. 26, no. 7, pp. 18699–18723, 2024, doi: 10.1007/s10668-023-03412-9.

[29] A. Lubis, Y. Irawan, Junadhi, and S. Defit, “Leveraging K-Nearest Neighbors with SMOTE and Boosting Techniques for Data Imbalance and Accuracy Improvement,” *J. Appl. Data Sci.*, vol. 5, no. 4, pp. 1625–1638, 2024, doi: 10.47738/jads.v5i4.343.

[30] K. Alnowaiser, “Improving Healthcare Prediction of Diabetic Patients Using KNN Imputed Features and Tri-Ensemble Model,” *IEEE Access*, vol. 12, no. 3, pp. 16783–16793, 2024, doi: 10.1109/ACCESS.2024.3359760.

[31] Z. Xu, C. Qiao, Y. Xiong, and J. Yu, “Enhancing Equipment Health Prediction with Enhanced SMOTE-KNN,” *J. Ind. Eng. Appl. Sci.*, vol. 2, no. 2, pp. 13–20, 2024.

[32] Herianto, B. Kurniawan, Z. H. Hartomi, Y. Irawan, and M. K. Anam, “Machine Learning Algorithm Optimization using Stacking Technique for Graduation Prediction,” *J. Appl. Data Sci.*, vol. 5, no. 3, pp. 1272–1285, 2024.

[33] H. Fonda, Y. Irawan, R. Melyanti, R. Wahyuni, and A. Muhaimin, “A Comprehensive Stacking Ensemble Approach for Stress Level Classification in Higher Education,” *J. Appl. Data Sci.*, vol. 5, no. 4, pp. 1701–1714, 2024.

[34] S. Rezaei Melal, M. Aminian, and S. M. Shekarian, “A machine learning method based on stacking heterogeneous ensemble learning for prediction of indoor humidity of greenhouse,” *J. Agric. Food Res.*, vol. 16, no. 2, pp. 1-11, 2024, doi: 10.1016/j.jafr.2024.101107.

[35] A. Zaidi, “Predicting wildfires in Algerian forests using machine learning models,” *Heliyon*, vol. 9, no. 7, pp. 1-13, 2023, doi: 10.1016/j.heliyon.2023.e18064.

[36] T. R. Mahesh, V. K. V, D. K. V, O. Geman, and M. Margala, "Healthcare Analytics The stratified K-folds cross-validation and class-balancing methods with high-performance ensemble classifiers for breast cancer classification," *Healthc. Anal.*, vol. 4, no. July, pp. 1–10, 2023.

[37] W. Kaleem, S. Tewari, M. Fogat, and D. A. Martyushev, "A hybrid machine learning approach based study of production forecasting and factors influencing the multiphase flow through surface chokes," *Petroleum*, vol 10, no. 2, pp. 354-371, 2023, doi: 10.1016/j.petlm.2023.06.001.

[38] M. Hasan, P. A. Bath, C. Marincowitz, L. Sutton, R. Pilbery, and F. Hopfgartner, "Pre-hospital prediction of adverse outcomes in patients with suspected COVID-19 : Development , application and comparison of machine learning and deep learning methods," *Comput. Biol. Med. J.*, vol. 151, no. May, pp. 1-10, 2022.

[39] S. Hou, Y. Liu, and Q. Yang, "Real-time prediction of rock mass classification based on TBM operation big data and stacking technique of ensemble learning," *J. Rock Mech. Geotech. Eng.*, vol. 14, no. 1, pp. 123–143, 2022

[40] R. Liu., "Stacking Ensemble Method for Gestational Diabetes Mellitus Prediction in Chinese Pregnant Women : A Prospective Cohort Study," vol. 2022, no. 1, pp. 1-12, 2022, doi: 10.1155/2022/8948082.

[41] H. Zheng, S. W. A. Sherazi, and J. Y. Lee, "A Stacking Ensemble Prediction Model for the Occurrences of Major Adverse Cardiovascular Events in Patients with Acute Coronary Syndrome on Imbalanced Data," *IEEE Access*, vol. 9, no. 2, pp. 113692–113704, 2021, doi: 10.1109/ACCESS.2021.3099795.

[42] S. S. Kshatri, D. Singh, B. Narain, S. Bhatia, M. T. Quasim, and G. R. Sinha, "An Empirical Analysis of Machine Learning Algorithms for Crime Prediction Using Stacked Generalization: An Ensemble Approach," *IEEE Access*, vol. 9, no. 3, pp. 67488–67500, 2021, doi: 10.1109/ACCESS.2021.3075140.