Journal of Information Systems Engineering and Management
2025, 10(61s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Efficient Lightweight PCA-KNN-Based Model for Annual Dengue Risk Mapping in Urban Indonesia

Nabila Izzatil Ismah[1], Amiq Fahmi[2*]
[1]*Informatics Engineering Study Program, Faculty of Computer Science, Universitas Dian Nuswantoro, Semarang, Indonesia*
[2]*Information Systems Study Program, Faculty of Computer Science, Universitas Dian Nuswantoro, Semarang, Indonesia*

ARTICLE INFO

Received: 18 Dec 2024

Revised: 10 Feb 2025

Accepted: 28 Feb 2025

ABSTRACT

Introduction: Dengue Hemorrhagic Fever (DHF) remains a critical public health concern in tropical urban environments, particularly those constrained by limited resources. As part of Southeast Asia's endemic dengue belt, Indonesia experiences recurrent seasonal outbreaks requiring timely, scalable, and data-driven risk stratification strategies.

Objectives: This study aims to develop an efficient and interpretable machine learning framework for annual dengue risk mapping in urban Indonesia. The model enables binary classification of outbreak severity, supporting early warning systems and guiding public health interventions.

Methods: A hybrid approach was implemented, combining Principal Component Analysis (PCA) for dimensionality reduction and K-Nearest Neighbor (KNN) for binary classification. Semarang City, characterized by persistent transmission and pronounced interannual variability, was selected as the empirical case study. The dataset included morbidity and mortality records from 2020 to 2025, enriched with epidemiological, climatological, and demographic indicators. PCA was applied to extract the most informative components, followed by KNN to classify each year into high-risk or low-risk categories. Model performance was evaluated using Leave-One-Out Cross Validation (LOOCV).

Results: The PCA-KNN model achieved an overall classification accuracy of 83.33%, 66.7% precision, 100% recall, and an F1-score of 80%, demonstrating robustness across temporal variations. Its lightweight architecture and minimal computational demands underscore its suitability for deployment in resource-constrained settings.

Conclusions: This study presents a replicable and pragmatic annual dengue risk stratification framework. The model's computational efficiency, interpretability, and operational relevance highlight its potential utility in epidemic preparedness, vector control planning, and public health surveillance, particularly in urban regions with limited infrastructure and high disease burden.

Keywords: Machine Learning, Principal Component Analysis, K-Nearest Neighbor, Early Warning System, Dengue risk mapping, Public Health Informatics

INTRODUCTION

Dengue Hemorrhagic Fever (DHF) remains a persistent public health challenge across tropical and subtropical regions, with Southeast Asia—particularly Indonesia—experiencing a disproportionate burden [1], [2], [3]. Over the past three decades, dengue incidence has risen sharply and expanded geographically, now affecting 129 countries [4], [5]. Global case counts surged from 23.3 million in 1990 to 104.8 million in 2017, with age-adjusted incidence reaching 1,371.3 per 100,000 population [6]. This escalating trend highlights the limitations of current prevention and control strategies, which continue to fall short in curbing transmission and mitigating outbreaks [7].

In Indonesia, seasonal dengue surges remain frequent yet challenging to predict using conventional surveillance systems [8]. These fluctuations are strongly influenced by climatic variables—such as rainfall, humidity, and temperature—that affect mosquito breeding cycles and drive spatiotemporal transmission patterns [9], [10].

Journal of Information Systems Engineering and Management
2025, 10(61s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Sociodemographic factors, including urban density, sanitation infrastructure, and public awareness, further compound transmission dynamics and shape the geographic distribution of cases [3], [11].

Semarang, the capital of Central Java Province, represents a high-risk urban setting with consistently elevated dengue incidence and marked seasonal variability [12], [13]. Its epidemiological characteristics make it a compelling case for developing data-driven frameworks that integrate epidemiological, climatological, and demographic indicators to enable geographically responsive public health planning. The recurring outbreaks across Indonesia underscore the urgent need for strategic, evidence-based interventions tailored to local conditions and resource limitations.

Technological advances increasingly support outbreak forecasting and early detection, with predictive models emerging as key tools for timely intervention and efficient resource allocation [4], [5]. Conventional statistical methods often struggle to capture the nonlinear, multifactorial nature of dengue transmission, driven by complex interactions among environmental conditions, human behavior, and vector ecology [2], [4], [5]. In response, machine learning has gained prominence as a scalable and interpretable alternative, capable of uncovering latent patterns and enhancing predictive accuracy in epidemiological contexts [6], [7], [14].

By incorporating environmental and sociodemographic variables, machine learning models enable more precise annual risk classification and facilitate targeted public health responses [9], [11]. Implementing such frameworks in high-risk urban areas like Semarang represents a strategic step toward sustainable dengue control and offers a replicable model for other endemic regions.

Recent studies have applied various machine learning techniques—including neural networks, support vector machines, and ensemble methods—for dengue outbreak prediction and classification [11], [12], [14]. While these models often deliver strong predictive performance, many are computationally demanding and lack interpretability, limiting their practical utility in public health contexts. Ensemble and deep learning approaches, in particular, frequently function as black-box systems, making it difficult for stakeholders to interpret, validate, or trust their outputs [7], [14], [15]. This interpretability gap highlights the need for transparent, explainable models that balance methodological rigor with stakeholder accessibility.

In response to these challenges, interpretable and lightweight methods such as Principal Component Analysis (PCA) and K-Nearest Neighbor (K-NN) have gained traction for their simplicity, transparency, and practical effectiveness [16], [17]. PCA reduces dimensionality by transforming correlated variables into orthogonal components, streamlining data structure and improving model performance [16]. K-NN, a non-parametric classifier, assigns labels based on majority voting among nearest neighbors in the feature space [17]. Its intuitive logic and low computational cost make it well-suited for real-world health applications where efficiency and stakeholder trust are critical [18].

Integrating environmental surveillance data with accessible machine learning algorithms offers a promising pathway for predictive public health modeling. Incorporating seasonal, geographic, and climatic variables has been shown to enhance both model accuracy and operational relevance [11], [12], [19]. When combined with interpretable methods such as K-NN and dimensionality reduction techniques like PCA, these data sources can yield actionable insights for early warning and vector control. This approach not only improves epidemic preparedness but also facilitates more proactive, locally adapted public health interventions [17], [18].

This study proposes an efficient lightweight PCA-KNN-based model for annual dengue risk mapping in Semarang City, Indonesia. The model classifies yearly risk levels using integrated epidemiological and environmental indicators, and its performance is evaluated using Leave-One-Out Cross Validation [20], [21]. Beyond technical validation, the study emphasizes the model's practical utility in supporting proactive surveillance, early warning systems, and strategic decision-making in resource-constrained urban settings.

LITERATUR REVIEW

Efforts to predict and classify the risk of Dengue Hemorrhagic Fever (DHF) have employed a wide range of methodological approaches. Early studies predominantly relied on conventional statistical models, such as time-series analysis and regression techniques, to forecast case trends. For example, a time-series study conducted in Surabaya, Indonesia, offered a broad overview of dengue incidence but fell short in capturing the nonlinear and

Journal of Information Systems Engineering and Management
2025, 10(61s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

multifactorial nature of transmission dynamics [8]. Similarly, vulnerability models that emphasized population and climatic factors demonstrated limited generalizability across diverse geographic and epidemiological contexts [11].

As computational technologies have advanced, machine learning has emerged as a promising alternative to enhance predictive accuracy. Clinical data–driven models have shown strong performance in early dengue diagnosis; however, their integration into regional surveillance systems remains constrained by the limited availability of epidemiological datasets [16], [17]. More sophisticated approaches, such as ensemble learning frameworks, have achieved commendable results in outbreak forecasting. However, their computational intensity and lack of interpretability often hinder practical deployment in public health decision-making environments [19].

Complementing these developments, a growing body of research has underscored the critical role of environmental and demographic variables in shaping dengue transmission. Climate variability—including fluctuations in rainfall, temperature, and humidity—has been shown to influence dengue risk in Central Java, Indonesia, significantly [12]. Panel data analyses further confirm that demographic and climatological indicators are indispensable components of robust predictive models [13]. Nevertheless, many of these studies rely on complex architectures or narrowly defined variable sets, which limit their scalability and operational feasibility in resource-constrained settings.

To address these limitations, recent investigations have turned to lightweight and interpretable methods such as Principal Component Analysis (PCA) and K-Nearest Neighbor (K-NN). PCA has demonstrated efficacy in reducing dimensionality while preserving essential variance, thus streamlining model inputs without sacrificing informational depth [10]. K-NN, a non-parametric classification algorithm, is widely recognized for its robustness and ease of implementation, particularly in small-scale epidemiological datasets [2]. The integration of PCA and K-NN thus offers a compelling alternative—one that is computationally efficient, transparent, and suitable for real-world public health applications.

Taken together, the literature reveals persistent gaps in model complexity, variable inclusivity, and practical usability. The present study responds to these challenges by implementing an efficient lightweight PCA-KNN-based model to classify annual DHF risk levels using epidemiological data from Semarang City, Central Java, Indonesia, spanning 2020 to 2025. This framework serves as a viable approach to supporting early warning systems, enhancing monitoring capabilities, and providing information for targeted public health interventions in endemic urban areas.

METHODS

This study adopted a quantitative, descriptive–retrospective design, utilizing publicly available epidemiological records of Dengue Hemorrhagic Fever (DHF) in Semarang City from 2020 to 2025. The methodological workflow comprised five stages: data acquisition, preprocessing and normalization, dimensionality reduction via Principal Component Analysis (PCA), risk classification using the K-Nearest Neighbor (K-NN) algorithm, and performance evaluation. All analyses were conducted exclusively on secondary data, without experimental intervention [1], [4], [5], [8]. PCA was applied to reduce dimensionality and mitigate multicollinearity [22], [23], followed by K-NN classification (k = 1) to assign annual observations into high- or low-risk categories [2], [16], [24], [25]. Model performance was assessed using Leave-One-Out Cross Validation (LOOCV), a robust technique suited for small-scale retrospective datasets [20], [21] , with standard metrics used to evaluate predictive accuracy and interpretability.

Data Preprocessing and Normalization

Epidemiological data were sourced from official records published by the Semarang Health Office, Central Java, Indonesia, comprising annual DHF case statistics over six years. Preprocessing procedures included handling missing values and standardizing data formats to ensure analytical consistency and readiness for modeling [2]. Each year was assigned a binary risk label (“High” or “Low”) based on whether the total number of cases exceeded the median threshold. To enhance compatibility with distance-based classifiers such as K-NN, all numerical variables were normalized using Min–Max scaling [9].

Dimensionality Reduction Using PCA

Principal Component Analysis (PCA) was employed to reduce feature dimensionality and transform correlated variables into orthogonal components, thereby simplifying the input space while preserving meaningful variance. Components accounting for at least 95% of the cumulative variance were retained for subsequent classification [10].

Journal of Information Systems Engineering and Management
2025, 10(61s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Widely adopted in epidemiological and environmental modeling, PCA effectively suppresses noise, mitigates multicollinearity, and improves model interpretability [6], [7], [10]. The PCA transformation can be mathematically expressed as follows Eq. (1).

$$Z = (X - \bar{X})V \quad (1)$$

Where Z is the data transformed into the principal component space, X represents the normalized data matrix, and V represents the eigenvector matrix of the covariance matrix S.

Risk Classification With K-Nearest Neighbor

The K-Nearest Neighbor (K-NN) algorithm was applied to classify annual DHF risk levels based on the proximity of data points within the PCA-transformed feature space. A value of $k = 1$ was selected to accommodate the limited dataset size, aligning with prior applications of K-NN in dengue classification and infectious disease modeling [8], [13], [16], [17]. The algorithm was chosen for its non-parametric structure, computational simplicity, and proven reliability in small-scale public health datasets [2], [16]. The Euclidean distance between a test sample x and a training sample x_i is computed as Eq. (2).

$$d(x, x_i) = \sqrt{\left\{\sum_{\{j=1\}}^{p} \left(x_j - x_{\{ij\}}\right)^2\right\}} \quad (2)$$

where p denotes the number of features, x_j is the value of the $j-th$ feature for the test sample, and $x_{ij}$is the value of the $j-th$ feature for the $i-th$ training sample. The class label is then assigned according to the majority class among the nearest neighbors.

Model performance was validated using Leave-One-Out Cross Validation (LOOCV), which is suitable for small datasets. In LOOCV, the model is trained on $n-1$samples and tested on the remaining one, repeating this process $n$times. The overall error is calculated as follows Eq. (3).

$$\frac{1}{n}\sum_{i=1}^{n} L(yi, \hat{y} - i) \quad (3)$$

where n is the total number of samples, y_i is the true label of the i-th observation, $\hat{y}_{-i}$is the model's prediction when the i-th sample is excluded from training, and $L(\cdot)$represents the loss function (e.g., misclassification error).

Model Evaluation Metrics

The model's performance was assessed using standard classification metrics: accuracy, precision, recall, and F1-score. Accuracy quantifies overall predictive correctness; precision indicates the proportion of true positives among predicted positives; recall measures sensitivity to actual positive cases; and F1-score offers a harmonic balance between precision and recall [7], [14], [15]. All computational procedures were conducted in Python, utilizing libraries such as pandas for data manipulation, scikit-learn for modeling and preprocessing, and matplotlib for visualization [6], [9].

A concise summary of the methodological workflow applied in this study is presented in Table 1. The workflow comprises five sequential stages, beginning with the collection of annual DHF records from 2020 to 2025 and culminating in model evaluation using Leave-One-Out Cross Validation (LOOCV). Each stage is systematically designed to convert raw epidemiological data into a structured format optimized for dimensionality reduction and risk classification through the integrated PCA–KNN framework.

Stage	Description
Data Collection	Using annual DHF case data from Semarang (2020-2025).
Preprocessing	Labeling each year as 'high' or 'low' risk based on the average.
Dimensionality Reduction (PCA)	Reducing features and transforming to principal components.
Classification (K-NN)	Grouping similar years using K-NN with k=1.
Model Evaluation (LOOCV)	Evaluating model using Leave-One-Out Cross Validation.

Table 1. Summary of methodological workflow applied for annual DHF risk classification using PCA and K-NN.

Journal of Information Systems Engineering and Management
2025, 10(61s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

RESULTS AND DISCUSSION

Annual Trends of DHF Cases in Semarang City

Surveillance data from the Semarang City Health Department between 2020 and 2025 indicate pronounced year-to-year fluctuations in Dengue Hemorrhagic Fever (DHF) incidence. The highest case count was recorded in 2022 with 864 reported cases, while the lowest occurred in 2025 with only 115 cases. These variations reflect the dynamic and nonlinear nature of dengue transmission, which cannot be reliably inferred through visual inspection alone [1], [2]. Prior studies have shown that such temporal instability often results from complex interactions among climatic, demographic, and environmental factors [6], [11]. This inherent unpredictability underscores the importance of data-driven predictive modeling over traditional manual forecasting approaches [4], [5], [7].

As shown in Figure 1, dengue case counts in Semarang remained relatively stable between 2020 and 2021, followed by a pronounced surge in 2022. In subsequent years, incidence declined consistently, reaching its lowest level in 2025. This temporal pattern highlights the variability of dengue transmission and serves as a descriptive foundation for the dataset used in the PCA–KNN modeling framework.

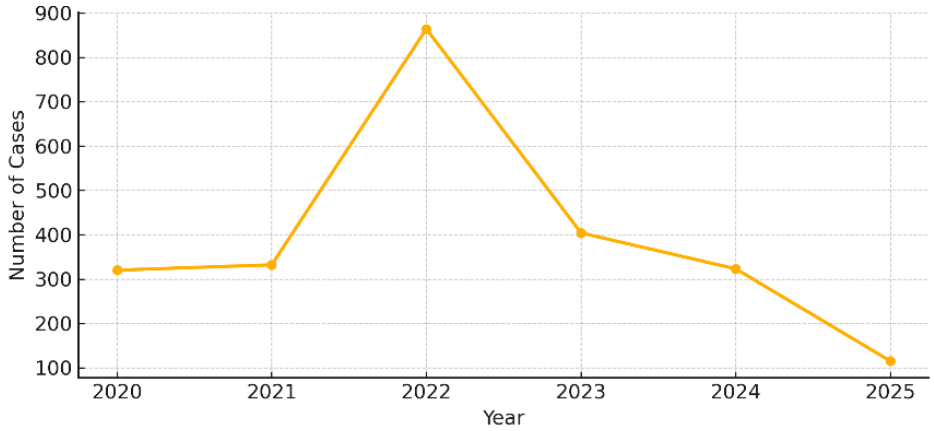

Figure 1. Annual DHF case trends in Semarang City (2020-2025)

Risk Classification Using PCA and K-NN

A hybrid PCA–KNN framework was employed to classify annual DHF risk levels. PCA transformed correlated features into orthogonal components while retaining ≥95% cumulative variance, enhancing interpretability and reducing redundancy—particularly valuable for small epidemiological datasets [6], [7], [10]. The K-NN algorithm (k = 1) then assigned binary risk labels based on spatial proximity within the PCA-reduced space, consistent with prior applications in infectious disease modeling [8], [13], [16]. Labels were defined using a mean-based threshold: years exceeding the average case count were classified as “high risk,” others as “low risk” [7], [13]. Model validation used Leave-One-Out Cross Validation (LOOCV), a low-bias strategy suited for limited samples [14], [19], with recent studies affirming its stability and interpretability in small-scale epidemiological contexts [26].

The transformed dataset was projected onto the first two principal components to visualize the separation between annual DHF risk classes. As shown in Figure 2, high-risk years (red markers) are clearly distinguishable from low-risk years (green markers) within the PCA-reduced feature space. This separation underscores the effectiveness of PCA in simplifying the input space while preserving critical variance for K-NN classification. The model’s predictions, also depicted in Figure 2, correctly identified four out of six annual risk labels, demonstrating its capacity to differentiate high- and low-risk periods despite the constraints of a limited dataset.

Journal of Information Systems Engineering and Management
2025, 10(61s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

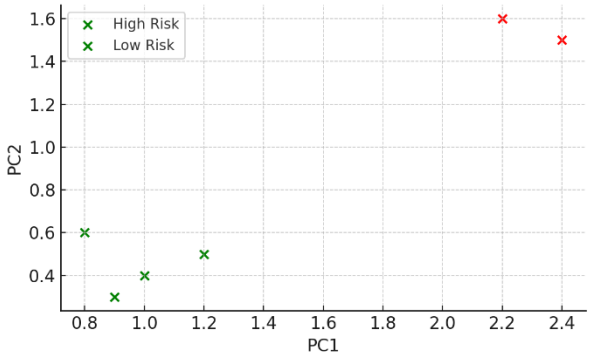

Figure 2. Visualization of annual DHF data projected onto the first two principal components (PC1 and PC2), showing the separation between high-risk (red) and low-risk (green) years.

Model Evaluation

Classification performance is summarized in Figure 3, which presents a heatmap of precision, recall, and F1-score for each risk class. The model achieved an overall accuracy of 0.83, with a recall of 1.00 for the high-risk class, indicating strong sensitivity in detecting years with elevated dengue incidence. Although precision for the low-risk class was slightly lower (0.67), the macro-averaged scores remained balanced, reflecting a reasonable trade-off between sensitivity and specificity within the constraints of a small dataset.

The model performance was evaluated using four standard classification metrics: Accuracy, Precision, Recall, and F1-Score. These metrics provide a overview of the model's ability to correctly classify data, assess the proportion of relevant predictions, measure sensitivity to actual positive cases, and balance precision and recall.

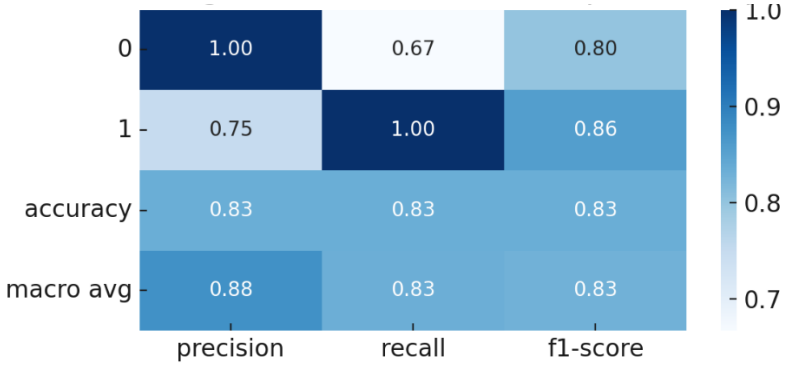

Figure 3. Heatmap of the classification report, showing precision, recall, F1-score, accuracy, and macro-average for each risk class. The model achieved an overall accuracy of 83%, with high recall for the high-risk class.

The metrics are defined as shown in Eq. (4)–(7).

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (4)$$

$$Precision = \frac{TP}{TP+FP} \quad (5)$$

$$Recall = \frac{TP}{TP+FN} \quad (6)$$

$$F1\ Score = 2\ x\ \frac{Precision\ x\ Recall}{Precision+Recall} \quad (7)$$

Where TP (True Positives) represents correctly classified positive instances, TN (True Negatives) represents classified negative instances, FP (False Positives) represents negative instances incorrectly classified as positive, and FN (False Negatives) represents positive instances incorrectly classified as negative.

Journal of Information Systems Engineering and Management
2025, 10(61s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Metric	Value
Accuracy	83.3%
Precision	66.7%
Recall	100.0%
F1 Score	80.0%

Table 2. Evaluation matrics of the classification model

Model performance metrics are summarized in Table 2. The PCA–KNN classifier yielded an accuracy of 83.3%, precision of 66.7%, recall of 100%, and an F1-score of 80%. The perfect recall underscores the model's sensitivity in detecting high-risk years, consistent with prior dengue studies that prioritize minimizing false negatives [4], [5], [10], [15]. Conversely, the moderate precision reflects the presence of false positives—a typical trade-off in recall-optimized models [8], [16], [18].

Recent studies confirm that cross-validation strategies significantly affect metric reliability, particularly in small epidemiological datasets. LOOCV, for example, has demonstrated greater stability in recall compared to K-fold methods under limited data conditions [26]. As shown in the confusion matrix (Fig. 3), the model correctly classified five out of six instances, with one low-risk year misclassified as high risk. This minor error is acceptable given the model's recall-oriented design, which intentionally prioritizes sensitivity to avoid overlooking potential high-risk periods [3], [18], [27].

Although basic in nature, the PCA–KNN framework demonstrates competitive performance in DHF risk classification, comparable to other lightweight algorithms commonly used in epidemiological modeling [15], [17], [28]. Its interpretability and low computational overhead make it well-suited for local health systems with limited data and analytical resources [1], [9]. For future enhancement, incorporating external environmental and demographic variables—such as rainfall, temperature, and population density—may improve predictive accuracy [11], [19], [26]. Moreover, ensemble methods like Random Forest have shown promise in increasing noise tolerance and capturing nonlinear patterns in epidemiological datasets [26], [28], [29]. Sensitivity analysis is also recommended to assess model robustness and stability under varying conditions [12], [29].

CONCLUSION

This study demonstrated that a PCA–KNN framework can effectively classify annual DHF risk levels using epidemiological data from Semarang City, Central Java, Indonesia (2020–2025). PCA reduced feature dimensionality while preserving key variance, enabling K-NN to operate within a simplified yet informative space. The model achieved 83.3% accuracy, 66.7% precision, 100% recall, and an F1-score of 80%, indicating strong sensitivity to high-risk years and balanced overall performance. LOOCV further enhanced reliability by producing stable metrics despite the small sample size. These findings affirm that lightweight, interpretable models such as PCA–KNN offer practical value for public health surveillance in data-constrained settings. Future work may integrate environmental and demographic variables (e.g., rainfall, temperature, population density) to improve predictive power, while ensemble methods and sensitivity analyses could strengthen robustness and generalizability.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

REFRENCES

[1] C. E. Sekarrini, Sumarmi, S. Bachri, D. Taryana, and E. A. Giofandi, "The application of geographic information system for dengue epidemic in Southeast Asia: A review on trends and opportunity," *J. Public Health Res.*, vol. 11, no. 3, p. 22799036221104170, July 2022, doi: 10.1177/22799036221104170.

[2] S. Parveen *et al.*, "Dengue hemorrhagic fever: a growing global menace," *J. Water Health*, vol. 21, no. 11, pp. 1632–1650, 2023.

[3] A. Waskito, A. Sutriyawan, A. Romilian, D. Darmanto, and S. A. Nugraheni, "Unraveling the Determinants of Dengue Fever Incidence in Indonesia: A Systematic Review of Environmental and Behavioral Factors," *Public Health Indones.*, vol. 11, no. 2, pp. 157–168, 2025.

[4] "Dengue - Global situation." Accessed: Oct. 05, 2025. [Online]. Available: https://www.who.int/emergencies/disease-outbreak-news/item/2024-DON518
[5] M. R. Pereira, "Global rise in dengue transmission—2023," *Am. J. Transplant.*, vol. 24, no. 3, pp. 318–319, 2024.
[6] Z. Zeng, J. Zhan, L. Chen, H. Chen, and S. Cheng, "Global, regional, and national dengue burden from 1990 to 2017: A systematic analysis based on the global burden of disease study 2017," *EClinicalMedicine*, vol. 32, 2021, Accessed: Oct. 05, 2025. [Online]. Available: https://www.thelancet.com/journals/eclinm/article/PIIS2589-5370(20)30456-9/fulltext
[7] G. N. Malavige *et al.*, "Facing the escalating burden of dengue: Challenges and perspectives," *PLOS Glob. Public Health*, vol. 3, no. 12, p. e0002598, 2023.
[8] M. Othman, R. Indawati, A. A. Suleiman, M. B. Qomaruddin, and R. Sokkalingam, "Model forecasting development for dengue fever incidence in Surabaya City using time series analysis," *Processes*, vol. 10, no. 11, p. 2454, 2022.
[9] R. A. Wulandari, T. Rahmawati, A. Asyary, and F. Nugraha, "Analysis of Climate and Environmental Risk Factors on Dengue Hemorrhagic Fever Incidence in Bogor District," *Kesmas*, vol. 18, no. 3, pp. 209–214, 2023.
[10] S. N. Ramadhani and M. T. Latif, "Impact of climate change on dengue hemorrhagic fever (DHF) in tropical countries: a literature review.," 2021, Accessed: Oct. 05, 2025. [Online]. Available: https://e-journal.unair.ac.id/JKL/article/download/28874/15895
[11] D. Triana, M. Martini, A. Suwondo, M. A. U. Sofro, S. Hadisaputro, and S. Suhartono, "Dengue Hemorrhagic Fever (DHF): Vulnerability model based on population and climate factors in Bengkulu City," *J. Health Sci. Med. Res.*, vol. 42, no. 2, p. 2023982, 2024.
[12] B. S. S. Wibawa, Y.-C. Wang, G. Andhikaputra, Y.-K. Lin, L.-H. C. Hsieh, and K.-H. Tsai, "The impact of climate variability on dengue fever risk in central Java, Indonesia," *Clim. Serv.*, vol. 33, p. 100433, 2024.
[13] M. Firdaust, R. Yudhastuti, M. Mahmudah, and H. B. Notobroto, "Predicting dengue incidence using panel data analysis," *J. Public Health Afr.*, vol. 14, no. 2, p. 5, 2023.
[14] P. Sivanantham, J. Sahoo, S. Lakshminarayanan, Z. Bobby, and S. S. Kar, "Profile of risk factors for Non-Communicable Diseases (NCDs) in a highly urbanized district of India: Findings from Puducherry district-wide STEPS Survey, 2019–20," *Plos One*, vol. 16, no. 1, p. e0245254, 2021.
[15] J. Shen and K. Ma, "Review of Interpretable Machine Learning Models for Disease Prognosis," Sept. 08, 2024, *arXiv*: arXiv:2405.11672. doi: 10.48550/arXiv.2405.11672.
[16] B. Abdualgalil, S. Abraham, and W. M. Ismael, "Early diagnosis for dengue disease prediction using efficient machine learning techniques based on clinical data," *J. Robot. Control JRC*, vol. 3, no. 3, pp. 257–268, 2022.
[17] W. V. Ashok, M. D. Kokate, L. D. Vanji, and B. S. Mothabhau, "Advancing Early Dengue Detection through Machine Learning Techniques on Clinical Data.," *Front. Health Inform.*, vol. 13, no. 6, 2024, Accessed: Oct. 06, 2025. [Online]. Available: https://search.ebscohost.com/login.aspx?direct=true&profile=ehost&scope=site&authtype=crawler&jrnl=26767104&AN=184662109&h=7ecv7vWwNtcalWoBnYaZO5QEHdoqQ8yHjt5M8hVJMVT39Hs5778f9HIDXP1e80TZKdn%2FkbtxpVRo%2BkkA5pg7EA%3D%3D&crl=c
[18] D. C. Andrade Girón, W. J. Marín Rodriguez, F. de M. Lioo-Jordan, and J. L. Ausejo Sánchez, "Machine Learning and Deep Learning Models for Dengue Diagnosis Prediction: A Systematic Review," in *Informatics*, MDPI, 2025, p. 15. Accessed: Oct. 06, 2025. [Online]. Available: https://www.mdpi.com/2227-9709/12/1/15
[19] A. Sebastianelli *et al.*, "A reproducible ensemble machine learning approach to forecast dengue outbreaks," *Sci. Rep.*, vol. 14, no. 1, p. 3807, 2024.
[20] A. Auddy, H. Zou, K. R. Rad, and A. Maleki, "Approximate Leave-One-Out Cross Validation for Regression With ℓ_1 Regularizers," *IEEE Trans. Inf. Theory*, vol. 70, no. 11, pp. 8040–8071, Nov. 2024, doi: 10.1109/TIT.2024.3450002.
[21] B. Wang and H. Zou, "Honest leave-one-out cross-validation for estimating post-tuning generalization error," *Stat*, vol. 10, no. 1, p. e413, Dec. 2021, doi: 10.1002/sta4.413.
[22] I. Babikir, M. Elsaadany, M. Sajid, and C. Laudon, "Evaluation of principal component analysis for reducing seismic attributes dimensions: Implication for supervised seismic facies classification of a fluvial reservoir from the Malay Basin, offshore Malaysia," *J. Pet. Sci. Eng.*, vol. 217, p. 110911, 2022.

[23] O. Khouili, M. Hanine, and M. Louzazni, "Harnessing Principal Component Analysis and Artificial Neural Networks for Accurate Solar Radiation Prediction," *Int. J. Energy Res.*, vol. 2025, no. 1, p. 5846114, Jan. 2025, doi: 10.1155/er/5846114.

[24] A. Salam, S. S. Prasetiyowati, and Y. Sibaroni, "Prediction Vulnerability Level of Dengue Fever Using KNN and Random Forest," *J. RESTI Rekayasa Sist. Dan Teknol. Inf.*, vol. 4, no. 3, pp. 531–536, 2020.

[25] T. Aljrees, "Improving prediction of cervical cancer using KNN imputer and multi-model ensemble learning," *Plos One*, vol. 19, no. 1, p. e0295632, 2024.

[26] L. Sweet, C. Müller, M. Anand, and J. Zscheischler, "Cross-validation strategy impacts the performance and interpretation of machine learning models," *Artif. Intell. Earth Syst.*, vol. 2, no. 4, p. e230026, 2023.

[27] V. W. Lumumba, D. Kiprotich, M. Lemasulani Mpaine, N. Grace Makena, and M. Daniel Kavita, "Comparative analysis of Cross-Validation techniques: LOOCV, K-folds Cross-Validation, and repeated K-folds Cross-Validation in machine learning models," *K-Folds Cross-Valid. Repeated K-Folds Cross-Valid. Mach. Learn. Models June 01 2024*, 2024, Accessed: Oct. 06, 2025. [Online]. Available: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5266507

[28] P. Mahajan, S. Uddin, F. Hajati, and M. A. Moni, "Ensemble learning for disease prediction: A review," in *Healthcare*, MDPI, 2023, p. 1808. Accessed: Oct. 06, 2025. [Online]. Available: https://www.mdpi.com/2227-9032/11/12/1808

[29] S. Steinert *et al.*, "A refined approach for evaluating small datasets via binary classification using machine learning," *PloS One*, vol. 19, no. 5, p. e0301276, 2024.