E-ISSN: 3025-5481
DINASTI INFORMATION AND TECHNOLOGY
https://dinastires.org/DIT dinasti.info@gmail.com +62 811 7404 455
DOI: https://doi.org/10.38035/dit.v4i1

Dengue Hemorrhagic Fever Case Prediction Using Climate and Social Media Data with Support Vector Regression

Erie Kresna Andana[1], Muhammad Nauvaldi Akbar A[2], Sri Amaliah Mandati[3]
[1]Department of Informatics, Faculty of Engineering, Universitas Muhammadiyah Surabaya, Indonesia, erie.kresna@um-surabaya.ac.id
[2]Department of Informatics, Faculty of Engineering, Universitas Muhammadiyah Surabaya, Indonesia, akbaralfaruqi58@gmail.com
[3]Department of Informatics, Faculty of Engineering, Universitas Muhammadiyah Surabaya, Indonesia, sri.amaliah.mandati@um-surabaya.ac.id

Corresponding Author: erie.kresna@um-surabaya.ac.id[1]

Abstract: Dengue hemorrhagic fever (DHF) remains a recurring public health problem in Surabaya, with transmission influenced by climatic variability and potentially reflected in social media activity. This study aims to develop an informatics-based model for predicting annual DHF cases by integrating climate and Twitter/X data and comparing Support Vector Regression (SVR) with Ridge Regression. The dataset covers 2010–2024 and combines annual DHF case records, annual and seasonal climate indices from Juanda and Tanjung Perak stations, and 320 Indonesian-language tweets related to DHF in Surabaya. Social media data were transformed into annual tweet volume, negative-sentiment tweet volume, and mean sentiment score, while predictors were aligned using a one-year time lag. Model performance was evaluated using Leave-One-Out Cross-Validation for 2011–2024 based on MAE, RMSE, and R^2. SVR achieved the best performance with an MAE of 386.31 cases, RMSE of 556.86 cases, and R^2 of 0.0121, outperforming Ridge Regression with an MAE of 601.31, RMSE of 727.19, and R^2 of −0.6846. These findings indicate that SVR better captures nonlinear relationships in multisource dengue data, although explanatory power remains limited by few annual observations and high feature dimensionality. This study demonstrates the potential of multisource data integration for data-driven dengue early warning systems.

Keywords: Dengue Hemorrhagic Fever, Support Vector Regression, Climate Data, Social Media, Machine Learning

INTRODUCTION

Dengue Hemorrhagic Fever is an arboviral disease transmitted primarily by *Aedes aegypti* mosquitoes and remains a serious public health concern in tropical and subtropical regions worldwide. The World Health Organization estimates that approximately 100–400 million dengue infections occur globally each year, while about 4.5 million cases were reported in the Region of the Americas alone in 2023, indicating that the global burden of dengue remains substantial (Zhang et al., 2025). In Surabaya, Dengue Hemorrhagic Fever also remains a recurring public health problem, with cases reported annually and patterns that are closely associated with environmental conditions and seasonal transitions. Variations in temperature, rainfall, and humidity can influence vector development and dengue transmission dynamics, making climatic information highly relevant for estimating changes in disease incidence. These conditions highlight the need for an informatics-based predictive approach capable of systematically processing historical data to support earlier risk detection and more effective disease-control planning.

Advances in information technology and the increasing availability of digital data have encouraged the use of data-driven and machine learning approaches for infectious disease surveillance and prediction. A systematic review by Sylvestre et al. (2022) demonstrated that the integration of real-world data, big data, and machine learning methods has considerable potential to improve dengue surveillance and forecasting, particularly when multiple sources of information are processed within an integrated framework. Fang et al. (2024) identified several significant climatic risk factors and demonstrated that machine learning algorithms can be applied to develop dengue outbreak prediction models based on meteorological characteristics. Mazhar et al. (2024) also reported that a Support Vector Machine-based approach achieved competitive performance in dengue outbreak classification using temperature, rainfall, humidity, and other meteorological variables. More recently, Tian et al. (2024) further reinforced the relevance of this approach by demonstrating that the integration of meteorological data into machine learning models can improve dengue prediction capability in Singapore.

In addition to meteorological information, the growth of social media has provided an alternative source of digital data that can dynamically reflect public responses, concerns, and perceptions regarding disease events. Public posts discussing symptoms, personal experiences, news, and concerns related to dengue can be processed as complementary digital signals alongside conventional health surveillance information. Babanejaddehaki et al. (2025) developed a Long Short-Term Memory-based approach combined with word embedding techniques to detect dengue- and influenza-related information from Twitter posts, demonstrating the potential of social media text as an epidemiological information source. In the Indonesian context, Anggraeni et al. (2024) combined an Indonesian-language Bidirectional Encoder Representations from Transformers model with a hybrid Convolutional Neural Network–Long Short-Term Memory architecture to classify dengue-related tweets and achieved a best accuracy of 0.91. These findings indicate that social media activity and sentiment characteristics have the potential to be transformed into predictive features that complement epidemiological and climatic data within an informatics-based prediction framework.

Despite the rapid development of dengue prediction research, an important research gap remains in how different data sources are integrated, as climate-based studies generally emphasize meteorological relationships with dengue incidence, whereas social media-based studies tend to focus on digital content detection, classification, or sentiment analysis. Consequently, the integration of official case records, climatic variables, and social media sentiment indicators within a single predictive framework still requires further investigation, particularly at the annual and city levels, such as in Surabaya. This gap is particularly relevant because the relationships between climatic conditions, public digital activity, and dengue incidence may not occur simultaneously, thereby requiring a time-lagged feature structure to represent potential historical effects on subsequent disease occurrence. Sebastianelli et al. (2024) demonstrated that an ensemble machine learning approach combining several base learners could produce more consistent dengue outbreak forecasts across multiple regions. However, their study used spatiotemporal data characteristics that differed from those of the present study, which relies on annual observations from a single city. Therefore, the scientific gap addressed in this study concerns not only the selection of a prediction algorithm but also the evaluation of whether integrated epidemiological, climatic, and social media data can provide useful predictive information under conditions of limited annual observations and relatively high predictor dimensionality.

Based on this research gap, the present study aims to develop a machine learning framework for predicting the annual number of Dengue Hemorrhagic Fever cases in Surabaya by integrating multiple data sources. The study uses Dengue Hemorrhagic Fever case data from 2010 to 2024, climate indices obtained from the Southeast Asia Climate Assessment and Dataset using observations from the Juanda and Tanjung Perak stations of the Indonesian Agency for Meteorology, Climatology, and Geophysics, and 320 Indonesian-language tweets related to dengue and Surabaya. Climatic information is represented through annual and seasonal indices, whereas social media data are transformed into total annual tweet volume, annual volume of negative-sentiment tweets, and annual mean sentiment score. All predictor variables are subsequently organized using a one-year time-lag scheme in which information from the preceding year is used to predict the number of Dengue Hemorrhagic Fever cases in the following year. The study then compares the predictive performance of Support Vector Regression as a nonlinear regression model with Ridge Regression as a regularized linear regression model using Leave-One-Out Cross-Validation.

Theoretically, this study contributes to the field of health informatics by developing a feature-integration framework that combines epidemiological, climatic, and social media data as predictors within a machine learning-based regression model. This approach extends previous dengue prediction research that has often relied on a single category of data by evaluating potential nonlinear relationships among climatic characteristics, public digital activity, and changes in disease incidence. Methodologically, the comparison between Support Vector Regression and Ridge Regression provides empirical evidence regarding the relative capability of nonlinear and regularized linear models when applied to annual datasets characterized by limited observations and relatively high feature dimensionality. From a practical perspective, the proposed framework can serve as an initial foundation for the development of a data-driven early warning system that may assist public health authorities in identifying changing dengue risks and planning more proactive control measures. Accordingly, this study contributes to the advancement of multisource data-driven surveillance while providing a context-specific application for annual Dengue Hemorrhagic Fever case prediction in Surabaya (K. I. O. Roster, 2022).

METHODS

Research Design and Data Sources

This study employed a computational quantitative approach with a predictive design based on secondary data from Surabaya for the 2010–2024 period. Three data sources were used: annual Dengue Hemorrhagic Fever case data, climate data, and social media data. The integration of epidemiological, climatic, and digital data is consistent with data-driven approaches increasingly applied to dengue surveillance and prediction (Chowdhury, 2026). Dengue case data for 2010–2014 were obtained from district-level records reported in Utami's undergraduate thesis, data for 2015 from Yuniarti's final project, and data for 2016–2024 from the Surabaya Health Profile published by the Surabaya City Health Office. Climate data were obtained from the *Southeast Asia Climate Assessment and Dataset* managed by the Indonesian Agency for Meteorology, Climatology, and Geophysics through the Tanjung Perak Maritime Meteorological Station and Juanda Meteorological Station, while the social media dataset consisted of 320 Indonesian-language tweets related to dengue and Surabaya collected from Twitter/X during 2010–2024.

Climate Variables and Data Preprocessing

Seven climate indices were used: daily mean temperature, daily minimum temperature, daily maximum temperature, total rainfall, number of rainy days, maximum consecutive dry days, and maximum consecutive wet days. The use of meteorological variables as predictors is supported by previous studies demonstrating their relevance in machine learning-based dengue prediction (K. Roster et al., 2022). Each index was calculated annually and for four seasonal periods: December–January–February, March–April–May, June–July–August, and September–October–November, producing 35 climate features. The tweets were cleaned and analyzed using a lexicon-based sentiment approach to generate three annual features: total tweet volume (TOTAL_TWEETS), negative-sentiment tweet volume (NEG_TWEETS_VOLUME), and mean sentiment score (MEAN_SENTIMENT), consistent with the use of social media data as supplementary information for dengue surveillance (Xiao et al., 2025). These features were integrated with annual dengue cases into a dataset of 15 years × 40 columns, consisting of one year variable, 38 predictor features, and one target variable.

Feature Engineering and Time-Lag Construction

A one-year time lag was applied to all predictor variables, where features from year t−1t-1 were used to predict dengue cases in year tt. Temporal characteristics are important in dengue prediction because climatic conditions may influence disease transmission after a certain time interval rather than simultaneously (Zaw et al., 2023). This procedure produced 14 observation pairs, with target years from 2011 to 2024 and predictor data from 2010 to 2023. The time-lag structure ensured that predictor information temporally preceded the target variable.

Model Development

Two regression models were developed and compared: Support Vector Regression with a Radial Basis Function kernel as the primary model and Ridge Regression as the regularized linear baseline model. All 38 predictor features were normalized before model training to reduce differences in measurement scales between climate and social media variables. The use of machine learning models is appropriate for dengue prediction because epidemiological and environmental predictors may exhibit complex and nonlinear relationships (Hussain et al., 2023). Previous research has also demonstrated the potential of machine learning approaches for forecasting dengue outbreaks using multiple predictor variables (Chen & Moraga, 2025).

Model Evaluation

Because the final dataset contained only 14 annual observations, model performance was evaluated using Leave-One-Out Cross-Validation. In each iteration, one observation was used as test data and the remaining 13 observations were used as training data until all observations had served as test data. Model performance was assessed using Mean Absolute Error, Root Mean Squared Error, and the coefficient of determination (R^2). The use of error-based metrics and the coefficient of determination provides complementary information regarding prediction accuracy and the model’s ability to represent variation in the target variable. The overall research workflow is summarized in Figure 1.

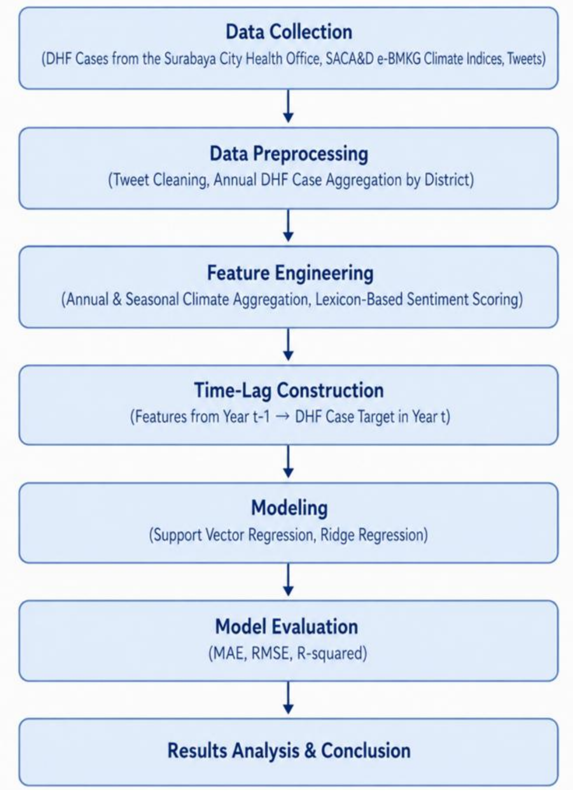

Figure 1. Research Flow Diagram

RESULTS

Integrated Dataset Characteristics

The integrated dataset covered the period from 2010 to 2024 and combined annual Dengue Hemorrhagic Fever case data, climate variables, and social media indicators. The descriptive variables presented in Table 1 include annual Dengue Hemorrhagic Fever cases, annual rainfall, annual mean temperature, total tweet volume, negative-sentiment tweet volume, and annual mean sentiment score. The number of Dengue Hemorrhagic Fever cases showed substantial variation across the observation period, ranging from 73 to 3,389 cases per year. The highest number of cases was recorded in 2010 with 3,389 cases, followed by 2,207 cases in 2013, whereas the lowest number was recorded in 2020 with 73 cases. After 2020, the number of cases increased gradually to 111 cases in 2021, 195 cases in 2022, 191 cases in 2023, and 231 cases in 2024.

Table 1. Summary of the Integrated Dataset, 2010–2024

Year	DHF Cases	Rainfall (mm)	Mean Temperature (°C)	Total Tweets	Negative Tweets	Mean Sentiment
2010	3389	2,787.9	28.29	1	1	-3.00
2011	1008	2,086.3	27.79	7	7	-1.43
2012	1091	1,376.7	28.07	4	3	-1.00
2013	2207	2,461.8	28.27	25	10	-0.32
2014	816	1,927.0	28.35	5	2	-0.20
2015	640	1,970.3	28.58	28	15	-0.57
2016	938	2,842.8	28.81	22	10	-0.09
2017	325	2,216.8	28.41	38	7	0.24

Year	DHF Cases	Rainfall (mm)	Mean Temperature (°C)	Total Tweets	Negative Tweets	Mean Sentiment
2018	321	1,740.8	28.45	34	12	-0.12
2019	277	1,629.1	28.55	51	13	0.20
2020	73	2,606.7	28.64	15	6	0.40
2021	111	2,435.8	28.45	20	10	-0.45
2022	195	2,592.5	28.11	38	9	0.50
2023	191	2,213.3	28.53	0	0	0.00
2024	231	1,834.1	28.95	0	0	0.00

Climate variables also varied across the study period. Annual rainfall ranged from 1,376.7 mm in 2012 to 2,842.8 mm in 2016, while annual mean temperature ranged from 27.79°C in 2011 to 28.95°C in 2024. These values indicate that the climate predictors exhibited temporal variation throughout the 15-year observation period. Such variability was retained in the integrated dataset and subsequently represented through annual and seasonal climate features in the prediction models.

The social media data also demonstrated substantial annual variation. Total tweet volume ranged from 0 to 51 tweets per year, with the highest volume recorded in 2019. The number of negative-sentiment tweets ranged from 0 to 15 tweets per year, with the highest value occurring in 2015. Mean annual sentiment scores ranged from −3.00 in 2010 to 0.50 in 2022, with predominantly negative values during the earlier observation years. No tweets meeting the study criteria were recorded in the integrated dataset for 2023 and 2024, resulting in total tweet volume, negative tweet volume, and mean sentiment values of zero for both years.

Overall, the integrated data demonstrate considerable temporal heterogeneity across epidemiological, climatic, and social media variables. The annual Dengue Hemorrhagic Fever case series contained both relatively high-incidence periods and periods with substantially fewer cases, while climate and social media indicators exhibited different temporal patterns. These characteristics formed the basis for constructing the 38 predictor features used in the prediction models.

4.2 Model Performance Evaluation

The predictive performance of Support Vector Regression and Ridge Regression was evaluated using Leave-One-Out Cross-Validation on the 14 target observations covering 2011–2024. Three evaluation metrics were used: Mean Absolute Error, Root Mean Squared Error, and the coefficient of determination. The results of the model evaluation are presented in Table 2.

Table 2. Comparison of Model Performance Based on Leave-One-Out Cross-Validation

Model	MAE (cases)	RMSE (cases)	R^2
SVR	386.31	556.86	0.0121
Ridge Regression	601.31	727.19	-0.6846

Support Vector Regression produced a Mean Absolute Error of 386.31 cases and a Root Mean Squared Error of 556.86 cases, whereas Ridge Regression produced a Mean Absolute Error of 601.31 cases and a Root Mean Squared Error of 727.19 cases. Thus, Support Vector Regression generated lower prediction errors than Ridge Regression for both error-based evaluation metrics. The coefficient of determination was 0.0121 for Support Vector Regression and −0.6846 for Ridge Regression. Based on all three evaluation metrics, Support Vector Regression provided the better predictive performance among the two models evaluated.

Although Support Vector Regression outperformed Ridge Regression, its coefficient of determination remained close to zero. This result indicates that only a limited proportion of the variation in annual Dengue Hemorrhagic Fever cases was represented by the fitted predictions. Ridge Regression produced a negative coefficient of determination, indicating poorer predictive performance relative to the observed variation in the target data. Therefore, the evaluation results identify Support Vector Regression as the better-performing model in this study, while also showing that the overall predictive capability remained limited.

4.3 Interpretation of Model Evaluation Results

The comparison of the two regression models indicates a consistent advantage of Support Vector Regression in terms of prediction error. Its lower Mean Absolute Error shows that, on average, the absolute difference between predicted and observed Dengue Hemorrhagic Fever cases was smaller than that produced by Ridge Regression. Similarly, the lower Root Mean Squared Error indicates that Support Vector Regression generated smaller overall deviations when larger prediction errors were given greater weight. These results are summarized in Table 3.

Table 3. Interpretation of Model Evaluation Results

Finding	Interpretation	Implication
SVR achieved lower MAE and RMSE values than Ridge Regression.	SVR was better able to model nonlinear relationships among climate factors, social media sentiment, and the number of DHF cases through its kernel function, whereas Ridge Regression assumes linear relationships.	DHF epidemiological data exhibit nonlinear characteristics, indicating that kernel-based models are more suitable than linear regression models.

The coefficient of determination provides additional information regarding model performance. The Support Vector Regression value of 0.0121 was higher than the Ridge Regression value of −0.6846, but remained close to zero. Accordingly, the results show that Support Vector Regression performed better in relative terms but did not fully capture the variability observed in annual Dengue Hemorrhagic Fever cases. The Ridge Regression result was weaker across all reported metrics.

The findings from Table 2 and Table 3 therefore distinguish two aspects of model performance. First, Support Vector Regression was the best-performing model among those evaluated because it consistently yielded lower Mean Absolute Error and Root Mean Squared Error and a higher coefficient of determination. Second, the absolute level of model performance remained limited, particularly as reflected by the low coefficient of determination. This distinction is important because the superiority of Support Vector Regression is relative to Ridge Regression and should not be interpreted as indicating near-perfect predictive accuracy.

4.4 Comparison Between Actual and Predicted Dengue Cases

The annual predictions generated by Support Vector Regression were compared with the actual number of Dengue Hemorrhagic Fever cases for the period 2011–2024, as illustrated in Figure 2. The predicted series generally followed the direction of several annual changes but showed substantially less variation than the actual case series. The largest discrepancy was observed during years characterized by relatively extreme case counts.

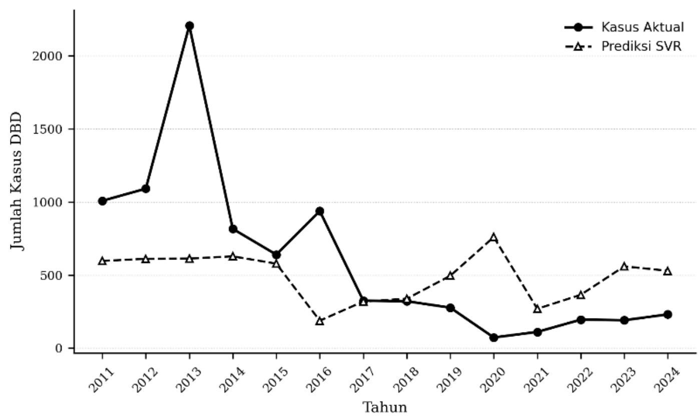

Figure 2. Comparison of Actual Dengue Hemorrhagic Fever Cases and Support Vector Regression Predictions (2011–2024)

In 2013, when the actual number of cases increased sharply to **2,207 cases**, the Support Vector Regression model produced a considerably lower predicted value, indicating underestimation of the observed peak. A similar limitation was visible for other periods in which actual case counts changed markedly between consecutive years. In contrast, during the low-incidence period of 2020–2022, the model generally generated predicted values above the corresponding actual case counts, indicating overestimation.

The prediction curve was therefore more concentrated around moderate case values than the actual case series. This pattern corresponds to a tendency toward *regression toward the mean*, in which predicted values remain closer to the central range of historical observations and do not reproduce extreme increases or decreases to the same extent as the observed data. Consequently, the Support Vector Regression model was more successful in representing the general level of annual cases than in reproducing extreme fluctuations.

4.5 Summary of the Main Findings

The results produced three principal findings. First, the integrated dataset exhibited substantial annual variation in Dengue Hemorrhagic Fever cases, climate indicators, and social media characteristics during 2010–2024. Second, Support Vector Regression consistently outperformed Ridge Regression, achieving a Mean Absolute Error of 386.31 cases, a Root Mean Squared Error of 556.86 cases, and a coefficient of determination of 0.0121, compared with 601.31 cases, 727.19 cases, and −0.6846, respectively, for Ridge Regression. Third, the comparison between actual and predicted cases demonstrated that the Support Vector Regression model was less responsive to extreme changes, particularly the high case count observed in 2013 and the relatively low case counts during 2020–2022.

Taken together, these findings directly address the study objective of evaluating annual Dengue Hemorrhagic Fever case prediction using integrated climate and social media predictors. The empirical results identify Support Vector Regression as the better-performing model among the two evaluated models, while the low coefficient of determination indicates that its predictive capability remains limited.

Discussion

Comparative Performance of the Prediction Models

The results demonstrate that Support Vector Regression (SVR) consistently outperformed Ridge Regression across all evaluation metrics. Support Vector Regression achieved a Mean Absolute Error (MAE) of 386.31 cases, a Root Mean Squared Error (RMSE) of 556.86 cases, and a coefficient of determination (R^2) of 0.0121, whereas Ridge Regression produced an MAE of 601.31 cases, an RMSE of 727.19 cases, and an R^2 of −0.6846. The lower prediction errors produced by Support Vector Regression indicate that a nonlinear modeling approach was relatively more suitable for representing the relationships among climatic variables, social media indicators, and annual Dengue Hemorrhagic Fever cases than the regularized linear approach used as the baseline. This interpretation is consistent with Devis et al. (2025), who demonstrated that support vector-based machine learning could effectively model dengue outbreaks using meteorological predictors. The result is also conceptually supported by Barboza et al. (2023), who found that climatic risk factors may interact with dengue occurrence in ways that can be captured through machine learning approaches rather than relying solely on conventional linear assumptions.

Nevertheless, the superiority of Support Vector Regression should be interpreted comparatively rather than as evidence of strong predictive accuracy. The R^2 value of 0.0121 is close to zero, indicating that the model explained only a very limited proportion of the observed variation in annual Dengue Hemorrhagic Fever cases. Furthermore, the negative R^2 value obtained by Ridge Regression indicates that its predictions were less effective than a simple reference prediction based on the mean of the observed target values. Therefore, the study supports the expectation that a nonlinear model performs better than the linear baseline under the current data structure, but it does not support a stronger assumption that the available multisource predictors are sufficient to explain annual dengue variation comprehensively.

Limited Sample Size and High Feature Dimensionality

One of the most important factors affecting model performance is the imbalance between the number of observations and the number of predictor variables. After applying the one-year time lag, the dataset contained only 14 annual observations but 38 predictor features, creating a high-dimensional setting relative to the available sample size. Under such conditions, a predictive model may identify relationships that are specific to the training observations but fail to generalize consistently to unseen data. Andrade Girón et al. (2025), in a systematic review of machine learning-based dengue prediction studies, found considerable heterogeneity in predictors, models, and predictive performance and emphasized that meteorological information is important but does not independently guarantee robust prediction.

The present findings also reinforce methodological concerns raised by Leung et al. (2023), who reviewed 99 dengue prediction models and identified limitations related to predictor selection, validation procedures, non-climatic data availability, and reporting quality. Consequently, the low R^2 obtained in the present study should not necessarily be interpreted as evidence that climate and social media variables are unrelated to dengue occurrence. Instead, it indicates that the available annual observations may be insufficient to estimate stable relationships among a relatively large number of predictors. The current results therefore highlight an important informatics issue: increasing the number of features does not automatically increase predictive performance when the amount of training data is highly constrained.

Temporal Resolution and Climate-Related Predictors

Another factor that may explain the limited model performance is the annual temporal resolution of the dataset. Dengue transmission is strongly influenced by short-term variations in rainfall, temperature, vector development, and environmental conditions, whereas annual aggregation compresses within-year fluctuations into a single observation. Tian et al. (2024) employed higher-resolution meteorological information in Singapore and demonstrated the potential of machine learning models to capture complex associations between meteorological variation and dengue occurrence. In comparison, the current study contains only one target observation for each year, which substantially reduces the number of temporal patterns available for model learning.

The use of a one-year time lag provides an important temporal structure because predictor information precedes the target cases, but the lag may also be relatively coarse for biological and epidemiological processes that occur over shorter periods. Del-Águila-Mejía et al. (2024) showed that dengue outbreak dynamics can be modeled by incorporating climate variability within a temporally dynamic epidemiological framework. Therefore, future research should evaluate monthly or weekly observations and compare several lag intervals rather than relying exclusively on a one-year lag. Such an approach would increase the number of observations and allow the model to identify whether rainfall, temperature, and other climate indicators provide predictive information at different temporal delays.

Role and Limitations of Social Media Data

The integration of Twitter/X data represents an important feature of this study because it introduces a digital population signal alongside epidemiological and climate data. However, the available volume was relatively small, ranging from 0 to 51 tweets per year, with only 320 Indonesian-language tweets across the entire 2010–2024 period. This sparsity may reduce the stability of annual sentiment indicators because a small number of posts can disproportionately influence total tweet volume, negative tweet volume, and mean sentiment. Salsabiila & Anggraeni (2025) demonstrated that dengue-related information can be extracted from Twitter using Long Short-Term Memory and word-embedding techniques, indicating the potential value of social media as a complementary epidemiological data source.

However, effective digital surveillance depends not only on the availability of social media content but also on adequate volume, filtering, and information extraction. (Konhar et al., 2026) emphasized the importance of systematic collection and filtration when processing social media streams for dengue surveillance because social media data are inherently noisy and unstructured. In the present dataset, the absence of identified tweets in 2023 and 2024 should therefore be interpreted cautiously; a value of zero in the analytical dataset does not necessarily imply the absence of public discussion about dengue, but rather the absence of tweets captured under the study's collection criteria.

The relevance of Indonesian-language social media data is further demonstrated by Anggraeni et al. (2024), who achieved an accuracy of 0.91 in classifying Indonesian dengue-related tweets using an Indonesian Bidirectional Encoder Representations from Transformers model combined with a hybrid Convolutional Neural Network–Long Short-Term Memory architecture. This finding suggests that richer textual representations could potentially extract more informative signals than the three aggregate annual indicators used in the present study. Nevertheless, advanced language models would also require substantially larger and appropriately labeled datasets, meaning that model complexity should be increased only after sufficient social media observations have been collected.

Multisource Data Integration and Scientific Contribution

The principal contribution of this study lies in the integration of epidemiological data, climate indices from the Southeast Asia Climate Assessment and Dataset, and Twitter/X sentiment indicators within a single predictive framework. This approach is consistent with the broader development of digital epidemiology, in which traditional surveillance data are supplemented with alternative digital data streams. Dettori et al. (2026) found that social media, internet-based information, climate variables, and other real-world data have considerable potential for dengue surveillance and prediction, while also emphasizing that these approaches should complement rather than replace established surveillance systems.

The integration of sentiment information also has a computational basis in previous research. Salehi et al. (2026) demonstrated that sentiment analysis can be combined with machine learning in dengue-related prediction and diagnostic modeling, illustrating the broader potential of textual and behavioral digital information. The present study extends this direction by integrating annual sentiment measures with climate and official epidemiological data at the city level. However, because the current design does not separately evaluate climate-only, social-media-only, and combined models, the specific incremental contribution of social media variables to predictive performance cannot yet be quantified.

Interpretation of Extreme Case Predictions

The comparison between actual and predicted values shows that Support Vector Regression was relatively better at representing the general level of annual cases than extreme fluctuations. The model underestimated the pronounced increase observed in 2013, while it tended to overestimate cases during the relatively low-incidence period of 2020–2022. This pattern is consistent with a regression-toward-the-mean tendency, in which predictions become concentrated around values frequently represented in the training data. With only 14 training-target observations, extreme epidemiological conditions are represented by very few examples, making it difficult for the model to learn their characteristics reliably.

This finding is important because accurate detection of unusually high incidence is particularly relevant for an early warning system. The results of Sebastianelli et al. (2024) demonstrate that ensemble machine learning can improve dengue forecasting when applied to substantially richer spatiotemporal datasets covering multiple geographical areas. Nevertheless, adopting a more complex ensemble architecture would not necessarily solve the present problem if the number of observations remains extremely small. Increasing data resolution and sample size should therefore precede or accompany increases in model complexity.

Implications for Data-Driven Dengue Surveillance

From an informatics perspective, the study provides evidence that heterogeneous information from different domains can technically be transformed and integrated into a unified machine learning dataset. The superior performance of Support Vector Regression relative to Ridge Regression also demonstrates the potential value of nonlinear modeling for epidemiological data in which climatic, behavioral, and disease-related variables may interact in complex ways. At the same time, the low coefficient of determination demonstrates that multisource integration alone does not guarantee predictive accuracy. The quality, quantity, temporal resolution, and representativeness of each source remain fundamental determinants of predictive performance.

Accordingly, the framework developed in this study should currently be considered a proof of concept for a data-driven dengue prediction framework rather than a deployment-ready early warning system. This distinction is important because operational public health decisions require models that demonstrate stable performance, appropriate validation, and reliability across different periods and populations. The present model can nevertheless provide a methodological foundation for future systems that integrate routine epidemiological surveillance with meteorological and digital information.

CONCLUSIONS AND RECOMMENDATIONS

Conclusion

This study developed a data-driven framework for predicting annual Dengue Hemorrhagic Fever cases in Surabaya by integrating epidemiological data, climate indices, and Twitter/X sentiment data for the 2010–2024 period using a one-year time-lag scheme. The predictive performance of Support Vector Regression was compared with Ridge Regression using Leave-One-Out Cross-Validation. Support Vector Regression achieved better performance, with a Mean Absolute Error of 386.31 cases, Root Mean Squared Error of 556.86 cases, and coefficient of determination of 0.0121, compared with Ridge Regression, which produced a Mean Absolute Error of 601.31 cases, Root Mean Squared Error of 727.19 cases, and coefficient of determination of −0.6846. These findings indicate that Support Vector

Regression was relatively more effective in capturing nonlinear relationships among climate variables, social media indicators, and annual dengue cases than the regularized linear model.

However, the low coefficient of determination indicates that the current model still has limited ability to explain the variation in annual Dengue Hemorrhagic Fever cases. This limitation is closely related to the small number of observations, consisting of only 14 annual target observations and 38 predictor features, as well as the limited availability of social media data. Nevertheless, the study demonstrates the feasibility of integrating heterogeneous epidemiological, climatic, and digital data within a machine learning framework. With improved data resolution, broader predictor coverage, and more robust modeling strategies, this approach has the potential to support the future development of a data-driven early warning system for dengue surveillance and public health decision-making.

Limitations

This study has several limitations that should be considered when interpreting the findings. First, the analysis was limited to Surabaya and annual observations from 2010 to 2024, which restricts the generalizability of the model to other geographical areas and limits its ability to capture short-term epidemiological and climatic variations. Second, after applying the one-year time lag, only 14 target observations were available for 38 predictor features, creating a high-dimensional setting that may reduce model stability and increase the risk of overfitting. Third, the social media dataset consisted of only 320 Indonesian-language tweets, with annual tweet volumes ranging from 0 to 51, which may not fully represent public responses to dengue events. In addition, the use of a fixed one-year time lag may not capture shorter temporal relationships between climatic conditions, digital activity, and dengue transmission. Therefore, the results should be interpreted within the methodological and data scope of this study.

Recommendations

Future studies should increase both the temporal and spatial resolution of the dataset by incorporating monthly or weekly observations, extending the study period, and including multiple geographical areas to provide a larger and more representative sample. The social media dataset should also be expanded to improve the reliability of digital surveillance indicators, while additional predictors such as entomological indices, population density, human mobility, socioeconomic conditions, land-use characteristics, and vector-control interventions may improve the explanatory and predictive capacity of the models. Future research may also apply feature-selection and dimensionality-reduction techniques, such as SHapley Additive exPlanations and Principal Component Analysis, to address the imbalance between the number of observations and predictor variables. Furthermore, broader comparisons involving ensemble learning, other machine learning algorithms, and different temporal validation strategies are recommended to identify models with stronger generalization performance and greater suitability for operational dengue early warning systems.

Author Contributions

All authors made substantial contributions to this study. The contributions included research conceptualization, development of the methodological framework, data collection and integration, data preprocessing, feature engineering, model development, statistical and computational analysis, interpretation of the findings, manuscript preparation, critical scholarly revision, and final approval of the manuscript. All authors have read and approved the final version of the article and agree to be accountable for the accuracy and integrity of the work.

Acknowledgment (Optional)

The authors acknowledge the Surabaya City Health Office and the Indonesian Agency for Meteorology, Climatology, and Geophysics for the availability of epidemiological and climate data used in this study. The authors also acknowledge the availability of Twitter/X data that supported the social media analysis conducted in this research. These data sources contributed to the development and evaluation of the multisource prediction framework presented in this study.

REFERENCES

Andrade Girón, D. C., Marín Rodriguez, W. J., Lioo-Jordan, F. de M., & Ausejo Sánchez, J. L. (2025). Machine learning and deep learning models for dengue diagnosis prediction: A systematic review. *Informatics*, *12*(1), 15.

Anggraeni, W., Kusuma, M. F. A., Riksakomara, E., Wibowo, R. P., & Sumpeno, S. (2024). Combination of BERT and Hybrid CNN-LSTM Models for Indonesia Dengue Tweets Classification. *International Journal of Intelligent Engineering & Systems*, *17*(1), 813.

Babanejaddehaki, G., An, A., & Papagelis, M. (2025). Disease outbreak detection and forecasting: A review of methods and data sources. *ACM Transactions on Computing for Healthcare*, *6*(2), 1–40.

Barboza, L. A., Chou-Chen, S.-W., Vásquez, P., García, Y. E., Calvo, J. G., Hidalgo, H. G., & Sanchez, F. (2023). Assessing dengue fever risk in Costa Rica by using climate variables and machine learning techniques. *PLOS Neglected Tropical Diseases*, *17*(1), e0011047.

Chen, X., & Moraga, P. (2025). Assessing dengue forecasting methods: a comparative study of statistical models and machine learning techniques in Rio de Janeiro, Brazil. *Tropical Medicine and Health*, *53*(1), 52.

Chowdhury, A. H. (2026). Data-Driven Machine Learning–Based Forecasting of Dengue in Bangladesh: Supporting Digital Health Approaches for Early Warning. *Health Science Reports*, *9*(3), e72147.

Del-Águila-Mejía, J., Morilla, F., & de Mata Donado-Campos, J. (2024). A system dynamics modelling and analytical framework for imported dengue outbreak surveillance and risk mapping. *Acta Tropica*, *257*, 107304.

Dettori, M., Deiana, G., Palmieri, A., Arghittu, A., Castiglia, P., Piana, A., & Campus, G. (2026). Rethinking Dengue Preparedness in the Era of Climate Change, Urbanisation, and Digital Health: A Structured Narrative Review. *Medicina*, *62*(7), 1333.

Devis, Y., Muhamadiah, M., Yulanda, Y., Irawan, Y., & Wahyuni, R. (2025). Optimization of machine learning models for risk prediction of DHF spread to support management strategies in urban areas. *Journal of Applied Data Sciences*, *6*(4), 2407–2420.

Fang, L., Hu, W., & Pan, G. (2024). Meteorological factors cannot be ignored in machine learning-based methods for predicting dengue, a systematic review. *International Journal of Biometeorology*, *68*(3), 401–410.

Hussain, Z., Khan, I. A., & Hassan, M. (2023). Machine learning approaches for dengue prediction: A review of algorithms and applications. *Pak. Geogr. Rev.*, *78*, 15–36.

Konhar, R., Lalsanga, J. K., & Biswal, D. K. (2026). Noise and neglect: Social-media signals expose attention gaps for dengue, chikungunya, lymphatic filariasis and kala-azar in India's vector-borne NTDs. *PLOS Neglected Tropical Diseases*, *20*(3), e0013378.

Leung, X. Y., Islam, R. M., Adhami, M., Ilic, D., McDonald, L., Palawaththa, S., Diug, B., Munshi, S. U., & Karim, M. N. (2023). A systematic review of dengue outbreak prediction models: Current scenario and future directions. *PLoS Neglected Tropical Diseases*, *17*(2), e0010631.

Mazhar, B., Ali, N. M., Manzoor, F., Khan, M. K., Nasir, M., & Ramzan, M. (2024). Development of data-driven machine learning models and their potential role in predicting dengue outbreak. *Journal of Vector Borne Diseases*, *61*(4), 503–514.

Roster, K., Connaughton, C., & Rodrigues, F. A. (2022). Machine-learning–based forecasting of dengue fever in Brazilian cities using epidemiologic and meteorological variables. *American Journal of Epidemiology*, *191*(10), 1803–1812.

Roster, K. I. O. (2022). *Data science for epidemiology: a case study of dengue in Brazil*. Universidade de São Paulo.

Salehi, M., Mousa Farkhani, E., Moghri, J., Ghasemian, A., Tabatabaee, S. S., & Hooshmand, E. (2026). Towards an integrated model for dengue management: a scoping review. *BMJ Open*, *16*(4), e111920.

Salsabiila, N. J., & Anggraeni, W. (2025). CNN BiGRU-Attention Classification and TiDE-PSO Forecasting Approach for Social Media-based Predictive Analysis of Dengue. *International Journal of Intelligent Engineering & Systems*, *18*(4), 716.

Sebastianelli, A., Spiller, D., Carmo, R., Wheeler, J., Nowakowski, A., Jacobson, L. V., Kim, D., Barlevi, H., Cordero, Z. E. R., & Colón-González, F. J. (2024). A reproducible ensemble machine learning approach to forecast dengue outbreaks. *Scientific Reports*, *14*(1), 3807.

Sylvestre, E., Joachim, C., Cecilia-Joseph, E., Bouzille, G., Campillo-Gimenez, B., Cuggia, M., & Cabie, A. (2022). Data-driven methods for dengue prediction and surveillance using real-world and Big Data: A systematic review. *PLoS Neglected Tropical Diseases*, *16*(1), e0010056.

Tian, N., Zheng, J.-X., Li, L.-H., Xue, J.-B., Xia, S., Lv, S., & Zhou, X.-N. (2024). Precision prediction for dengue fever in Singapore: A machine learning approach incorporating meteorological data. *Tropical Medicine and Infectious Disease*, *9*(4), 72.

Xiao, Y., Soares, G., Bastos, L., Izbicki, R., & Moraga, P. (2025). Dengue nowcasting in Brazil by combining official surveillance data and Google Trends information. *PLOS Neglected Tropical Diseases*, *19*(8), e0012501.

Zaw, W., Lin, Z., Ko Ko, J., Rotejanaprasert, C., Pantanilla, N., Ebener, S., & Maude, R. J. (2023). Dengue in Myanmar: Spatiotemporal epidemiology, association with climate and short-term prediction. *PLOS Neglected Tropical Diseases*, *17*(6), e0011331.

Zhang, W.-X., Zhao, T.-Y., Wang, C.-C., He, Y., Lu, H.-Z., Zhang, H.-T., Wang, L.-M., Zhang, M., Li, C.-X., & Deng, S.-Q. (2025). Assessing the global dengue burden: Incidence, mortality, and disability trends over three decades. *PLoS Neglected Tropical Diseases*, *19*(3), e0012932.