Journal of Artificial Intelligence Practice (2026)
Clausius Scientific Press, Canada

DOI: 10.23977/jaip.2026.090101
ISSN 2371-8412 Vol. 9 Num. 1

Exploring an Effective Machine Learning Method for Dengue Fever Prediction

Weifeng Wang

Wuhan Britain-China School, Wuhan, Hubei, China
wangweifeng0316@foxmail.com

Keywords: Dengue, Prediction, Public Health, Modeling, Machine Learning

Abstract: This study aims to build models based on the spread of dengue fever to predict its epidemic trends in different regions. Dengue fever is a mosquito-borne disease. Climate change, such as temperature and precipitation, is closely related to its spread, which is a major concern for public health in recent years. Taking the cities of San Juan and Iquitos as examples, this study uses machine learning to predict the trend. The model development tried methods such as random forest regression, KNN, XGBoost, LSTM, and support vector regression. The XGBoost performed best for San Juan while SVR excelled for Iquitos

1. Introduction

Dengue fever is a mosquito-borne disease that has posed a serious threat to global public health in recent years. Studies have shown that prediction models based on machine learning can improve the efficiency of public health prevention and control and provide a reference for the accuracy of different methods in predicting infectious diseases.

1.1 Background Information

Traditional methods for predicting dengue fever rely mostly on statistical models and time series models - autoregressive moving averages and Markov chains - however, such methods often assume a linear relationship between variables and are mainly based on basic meteorological data to analyze and fail to reveal the complex nonlinear dynamics of dengue fever transmission. Although it has certain effects, due to the spatiotemporal dynamic characteristics of dengue fever such as spatial spread and seasonal fluctuations, it is often slow to respond and difficult to deal with, hindering the accuracy of predictions and bringing disadvantages to public health decision-making.

In recent years, prediction models based on machine learning have been gradually used in dengue fever epidemiological research. Machine learning can capture the relationships between multiple factors through complex nonlinear mapping and integrate multi-source data such as meteorology, environment and socio-economics to significantly improve prediction accuracy. Among them, models such as support vector machines and long short-term memory networks perform well in dengue fever prediction. However, machine learning problems also include the over-fitting problem of the model, which fails to capture the general laws of the data, and the black-box training process that is difficult to control.

This study aims to use multi-source data fusion and feature engineering to build an efficient dengue fever prediction model to overcome the limitations of traditional methods and more accurately characterize the nonlinear relationship and spatiotemporal dynamic changes of dengue fever spread. This innovative method not only provides timely epidemic warning information, but also provides important support for public health emergency response and risk assessment.

1.2 Hypothesis

Newly explored machine learning models with advanced feature engineering perform better than traditional statistical and machine learning models, which will be more accurate for dengue prediction, even more when combined with consistent dengue seasonality data features.

2. Literature Review

The driving role of climate variables on dengue epidemic dynamics is a core theme of related research. Lowe et al.[1], based on a generalized linear mixed model (GLMM) study in southeastern Brazil, found that seasonal climate variables such as temperature and precipitation can significantly enhance the spatial and temporal resolution capabilities of dengue prediction models. In addition, studies have shown that large climate patterns such as El Niño have a non-negligible indirect impact on epidemic dynamics. In Guangzhou, China, Xu et al.[2] further demonstrated the time-lag effects of precipitation and temperature on mosquito vector density and virus transmission rate, thus revealing the intrinsic link between climate driving and seasonal outbreaks of dengue fever.

Traditional statistical models have shown some advantages in capturing the association between climate variables and the number of dengue cases. Hii et al.[3] analyzed the lagged effects of temperature and precipitation based on Poisson regression and successfully predicted the 16-week forward trend of the dengue fever epidemic in Singapore. Similarly, Gharbi et al.[4] achieved 3-month advance epidemic prediction in the Guadeloupe region through the SARIMA model. However, this type of method assumes that the relationship between variables is linear, and the predictive ability of the model relies on a relatively simple feature set, making it difficult to effectively handle high-dimensional data and nonlinear interactive features.

With the expansion of data scale and improvement of computing power, the application of machine learning methods in dengue fever prediction has gradually attracted attention. A systematic review by Leung et al.[5] pointed out that machine learning algorithms such as support vector machines (SVM) and long short-term memory networks (LSTM) have significant advantages in capturing the dynamic characteristics of time series. In addition, ensemble learning models (such as XGBoost) outperform traditional methods in variable selection and prediction accuracy. Recent comparative studies have further validated these findings. Chen and Moraga[7] conducted a comprehensive assessment of dengue forecasting methods in Rio de Janeiro, comparing statistical models with machine learning techniques and demonstrating the superior performance of advanced algorithms. Deep learning approaches have shown particular promise in various regional contexts. Bui et al.[9] developed deep learning models for dengue forecasting based on climate data in Vietnam, while Phan et al.[8] leveraged climate data with advanced machine learning approaches in Ba Ria Vung Tau Province. Furthermore, Moreira et al.[10] demonstrated that novel feature selection approaches combined with meteorological variables can significantly improve dengue forecasting accuracy in Bangladesh.

In summary, the combination of climate-driven variables with statistical and machine learning methods provides theoretical support and technical tools for dengue fever prediction. Based on the above literature, this study will further explore efficient machine learning models through multi-source data fusion and optimization of feature engineering to overcome the limitations of traditional methods and improve the model's adaptability to nonlinear characteristics and regional dynamics.

3. Experiment Methodology

3.1 Data Overview and Preprocessing

The dataset used in this study contains weekly environmental variables and the corresponding number of dengue cases in San Juan, Puerto Rico and Iquitos, Peru, while the test set provides environmental variables in the same time period for predicting the number of future cases. These data have a long time span and can capture seasonal and interannual changes, providing rich time series information for research and are provided by the Centers for Disease Control and Prevention and the National Oceanic and Atmospheric Administration, which are highly authoritative in the fields of environment and epidemiology. The dataset covers a variety of climate-related variables, such as environmental variables such as temperature, precipitation, and vegetation index, with a long time range and is recorded weekly with a high resolution. The reason for choosing these data is that the spread of dengue is closely related to climate change, and environmental variable data can well reflect this relationship. Additionally, the dataset focuses on high-incidence dengue regions and the research results are of practical significance. High-quality data consistent with scientific literature and expert consensus provide a strong basis for supporting model development and improving prediction accuracy.

The San Juan dataset spans from April 1990 to June 2008 (n=936 weeks), while the Iquitos dataset covers from July 2000 to June 2010 (n=520 weeks). Each dataset includes weekly dengue case counts and 20 environmental variables.

The data was first cleaned and missing values were filled by statistical methods. Then, the range of environmental variables in the training set was trimmed based on the maximum and minimum values of the test set to ensure the consistency of feature ranges. In addition, lagged variables were introduced through feature engineering to capture time dependencies, and all environmental variables were standardized to improve the training efficiency and performance of the model. Finally, outliers were processed to remove noise data that may affect model performance, ensuring the structured, consistent and high-quality data, laying a solid foundation for subsequent modeling work.

The dataset was split into training (70%) and testing (30%) sets using temporal ordering to preserve time series structure. Cross-validation was performed using 5-fold time series split for hyperparameter tuning.

3.2 Exploratory Data Analysis

After first drawing a line chart of the time series number of cases in the two cities of San Juan and Iquitos, the time series changing characteristics of dengue fever cases in the two cities are clearly presented (see Figure 1). The number of cases in San Juan fluctuates significantly and shows strong seasonal changes, especially at the end of each year and the beginning of the year, when the number of cases often reaches a significant peak. This trend may be influenced by local climate characteristics, such as seasonal changes in temperature and precipitation. However, the number of cases in Iquitos changes relatively steadily and lacks a significant seasonal high incidence pattern, indicating that the spread of dengue fever may be affected by stable climate conditions or dominated by different transmission mechanisms.

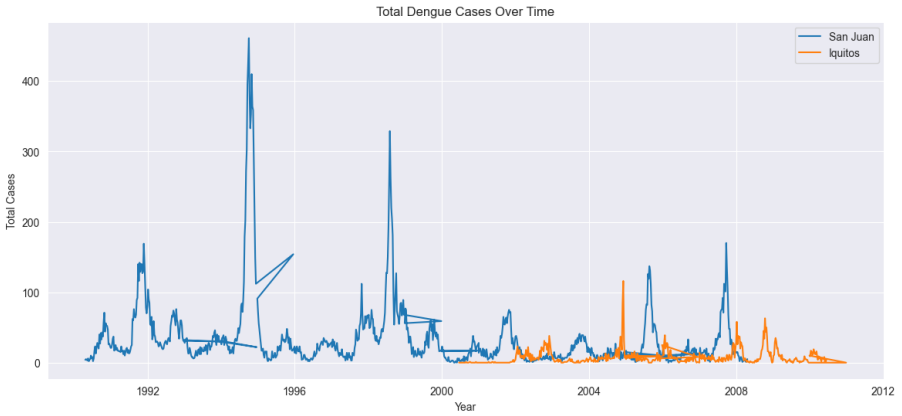

Figure 1. Time Series Plot of Dengue Cases over Time for San Juan and Iquitos

The Pearson correlation coefficient between the numerical features and the target variable total cases was further calculated, revealing the linear relationship between different climate variables and the number of cases (Figure 2). Humidity-related features, such as analysis_specific_humidity_g_per_kg and reanalysis_dew_point_temp_k, show a strong positive correlation with the number of cases, indicating that humid climate conditions may provide an ideal environment for mosquito reproduction and virus transmission. The correlation of NDVI-related features (such as ndvi_ne and ndvi_nw) is relatively weak, which may imply that the impact of vegetation cover on dengue fever transmission is indirect or limited to specific conditions.

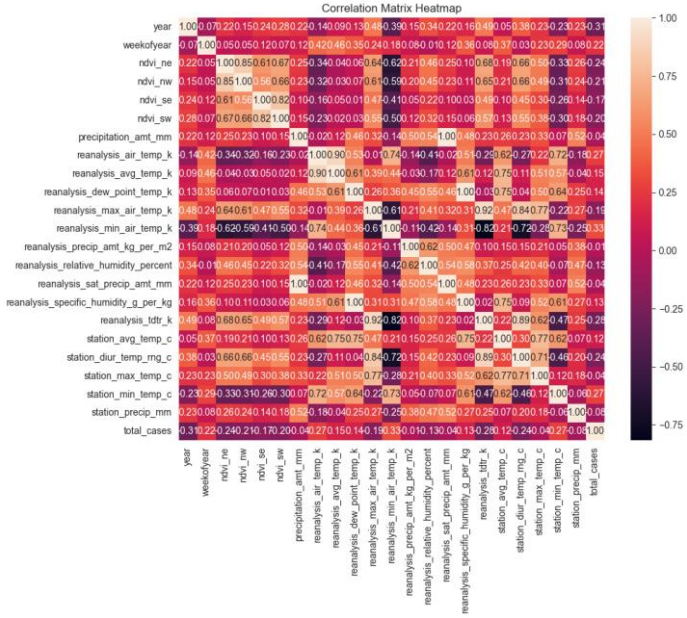

Figure 2. Correlation Matrix Heatmap of Climate Variables with Dengue Cases

3.3 Feature Engineering

In order to effectively capture the dynamic characteristics of time series and the potential impact of climate factors on the spread of dengue fever, this study conducted in-depth modeling and screening of multi-dimensional data through refined feature engineering strategies. The introduction of lag features, the construction of seasonal variables, and the normalization of numerical features not only greatly improve the model's ability to capture time dependence and seasonal fluctuations, but also further optimize feature selection through correlation analysis to ensure that features have

dimensional validity and model generalization ability.

The design of time lag features constructs 1-, 2-, and 4-week lag features based on variables such as the total number of cases, precipitation, relative humidity, and average temperature. These features bring past variable values into the current prediction time point through a translation operation along the time axis, thereby capturing the changing trends in cases caused by fluctuations in environmental conditions. In order to avoid the impact of missing values caused by lag operations, a linear interpolation method was used to fill the generated data. Experimental verification shows that after introducing the lagging feature, the model's ability to capture the surge and sudden drop in cases is significantly enhanced, especially the prediction accuracy for cyclical peaks and troughs is significantly improved.

In terms of constructing seasonal variables, seasonal division based on weekly series (weekofyear) provides key period information for the model. In the San Juan data, based on the time characteristics of dengue fever transmission, each year is divided into three major stages: slowing season, trough season, and climbing season, and the corresponding seasonal indicator variables are generated through binarization. This processing eliminates temporal inconsistencies in data across years by dynamically adjusting for week-day offsets. In the Iquitos data, it is further refined into the four seasons of autumn, winter, spring, and summer, and corresponding seasonal binary variables are generated. These seasonal features effectively capture the seasonal fluctuations in dengue transmission, manifesting as significant improvements in cyclical predictions in model validation.

In order to reduce the interference of different dimensions and scales in the data on model training, all numerical features are normalized. Specifically, MinMaxScaler is used to compress the feature values to the [0, 1] interval, thus ensuring the equal weight between features and optimizing the convergence speed of the gradient descent algorithm. The normalized data not only improves the performance of the activation function in the deep learning model, but also significantly reduces the convergence fluctuations caused by numerical imbalance during the training process.

By analyzing the heat map (Figure 2), it was found that certain climate variables are significantly correlated with the total number of cases, indicating that they have an important impact on the spread of dengue fever. For example, the correlation coefficient between relative humidity (reanalysis_relative_humidity_percent), showing the important role of humidity conditions in mosquito survival and virus replication. In addition, the vegetation index in the northeast (ndvi_ne) and southwest (ndvi_sw) regions showed positive correlations, indicating that vegetation coverage has an important impact on mosquito habitat conditions. To avoid multicollinearity problems between features, for variable pairs with a correlation coefficient higher than 0.9, only the features with more representative information expression are retained. For example, reanalysis_relative_humidity_percent was selected among the humidity variables, and station_avg_temp_c among the temperature variables to maximize the computational efficiency and prediction performance of the model.

Table 1 summarizes the definitions of core features and their potential contribution to dengue case prediction.

Table 1: Definition of core features and their potential impact

Feature Name	Definition	Potential Impact
Lagging characteristics of the total number of cases	Lag the total number of cases by 1 week, 2 weeks, and 4 weeks	Capture the dynamic trend of short-term cases to effectively respond to surges or sudden drops

Precipitation lag characteristics	Weekly precipitation totals lagged by 1 week, 2 weeks, and 4 weeks	Simulate changes in humidity and water accumulation conditions to reveal the dengue fever transmission environment
Relative humidity lag characteristics	Relative humidity lagged by 1 week, 2 weeks, and 4 weeks	Provide dynamic change information on mosquito habitats
Temperature lag characteristics	Temperature averages lagged by 1 week, 2 weeks, and 4 weeks	Capture the cyclical impact of climate conditions on virus transmission
Seasonal binary features	Seasonal indicator variables based on time periods	Provide periodic background information on the spread of dengue fever
Vegetation Index (NDVI)	Regional quantitative index of vegetation coverage level	Indirectly reflect the environmental characteristics of mosquito habitats

3.4 Model Implementation

In the data modeling process, a variety of machine learning and deep learning models are used, from traditional linear models to complex nonlinear models, covering the tree-based enhanced gradient boosting model XGBoost, linear regression/random forest, kernel-based support vector machine methods, as well as deep learning models such as multi-layer perceptron, autoencoder, convolutional neural network, long short-term memory network, and gated recurrent unit, as summarized in Table 2.

Table 2. Summary of Models and Key Characteristics

Model Type	Optimization Strategies	Key Characteristics
XGBoost	max_depth, eta, subsample	Efficient, nonlinear modeling
Random Forest	n_estimators, max_depth	Stable, interpretable
SVR	Kernel type, C, epsilon	Handles complex nonlinearity
KNN	n_neighbors, distance metric	Sensitive to neighbors
MLP	Hidden layers, Dropout	High-dimensional modeling
Autoencoder	encoding_dim, learning rate	Dimensionality reduction
CNN	num_filters, pooling	Extracts local features
LSTM / GRU	Hidden units, Dropout	Captures long dependencies

In terms of selecting and optimizing the tree model, XGBoost was optimized using a systematic grid search on core hyperparameters, including maximum depth, learning rate, subsampling ratio, and feature column sampling ratio. Under each set of parameter configurations, the performance of the model on the training set and validation set was evaluated through multiple cross-validations, and the parameter combination that performed best on the validation set was finally selected. Multiple evaluation indicators, such as mean absolute error (MAE) and coefficient of determination (R^2), were used to verify the generalization ability of the model.

Random Forest was optimized by adjusting key parameters such as number of trees, maximum

depth, minimum number of sample splits, minimum number of leaf node samples, and feature sampling ratio. Support Vector Regression (SVR) was tuned through core parameters, including kernel function type, penalty coefficient, and slack variable. The K-nearest Neighbor model was optimized by adjusting number of neighbors, distance measures, weighting strategies, and search algorithms.

Neural network architectures were optimized systematically. For Multi-Layer Perceptrons (MLP), key hyperparameters such as number of hidden layers, number of neurons in each layer, Dropout ratio, and learning rate were tuned. Autoencoders were explored for feature dimensionality reduction by adjusting dimensionality of the latent representation and learning rate. Convolutional Neural Networks (CNN) were tuned for number of convolution kernels, size of the convolution kernel, and pooling strategies. Long Short-Term Memory networks (LSTM) and Gated Recurrent Units (GRU) were optimized for number of hidden units, number of network layers, Dropout ratio, and learning rate.

Hyperparameter optimization was conducted using grid search with the following ranges:

- XGBoost: learning_rate [0.01, 0.1, 0.3], max_depth [3, 5, 7], n_estimators [50, 100, 200]
- SVR: C [0.1, 1, 10], gamma [0.001, 0.01, 0.1]
- Random Forest: n_estimators [50, 100, 200], max_depth [10, 20, 30]

3.5 Evaluation Metrics

First, the mean absolute error (MAE) was used to evaluate the average deviation between the predicted value and the true value, reflecting the overall level of model prediction accuracy. Second, the coefficient of determination (R^2) was used to measure the model's ability to explain the variance of the target variable, reflecting the model's fitting effect from a global perspective. In addition, three-dimensional residual graph was drawn, using time, the number of true cases, and the residuals as dimensions to intuitively display the possible systematic errors and time-related trends in the model's predictions, providing reliable support for the comprehensive evaluation of the model's performance.

4. Results and Evaluation

In Table 3, the MAE and R^2 performance data of each model for the two regions of San Juan and Iquitos are summarized to provide an overall perspective for subsequent analysis.

Table 3. Summary of model performance in San Juan and Iquitos

Region	Model	MAE	R^2
San Juan	XGBoost	22.8427	0.3077
	MLP	23.1595	0.2713
	LSTM	23.3562	0.1860
	SVR	23.5607	0.0655
	Random Forest	23.5708	0.2637
	GRU	24.3646	0.1843
	CNN	24.4719	0.1626
	Autoencoder	24.9328	0.1653
	Linear	25.4742	0.1475

	KNN	27.2869	0.0471
Iquitos	SVR	6.0932	0.1278
	LSTM	6.1573	0.1334
	CNN	6.2873	0.0222
	Autoencoder	6.3586	0.0420
	Linear	6.4101	0.0424
	XGBoost	6.4403	0.0833
	GRU	6.5201	0.1458
	MLP	6.5770	0.1151
	Random Forest	6.7002	0.0748
	KNN	6.7243	0.0317

4.1 Model Performance Evaluation

In the dengue fever prediction in San Juan, the performance comparison of different machine learning models is shown in Figure 3. The XGBoost model performed best, with a mean absolute error (MAE) of 22.84 and a coefficient of determination (R^2) of 0.3077, indicating that the model can better capture the nonlinear relationship of the data. In contrast, the MLP model performed second, with a MAE of 23.16 and an R^2 of 0.2713, demonstrating its ability to model high-dimensional features. The MAEs of LSTM and GRU were 23.36 and 24.36, respectively, but their R^2 values were low, only 0.1860 and 0.1843, indicating that they failed to fully exploit the time series features when processing San Juan data. The KNN model has the worst performance, with a MAE of 27.29 and an R^2 of only 0.0471, and it is difficult to handle high-dimensional feature data. This shows that tree-based models (such as XGBoost) and neural network models perform better than traditional linear and KNN models on the San Juan dataset.

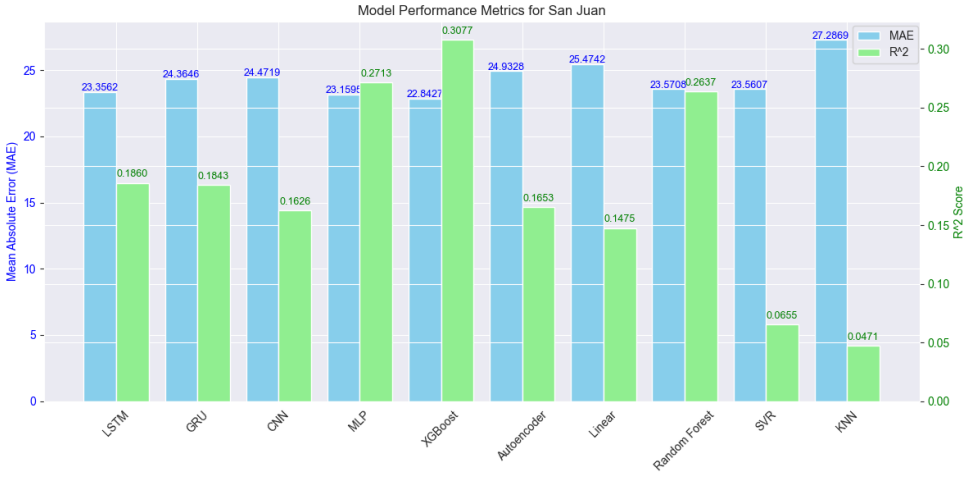

Figure 3. Comparison of MAE and R^2 of various models in San Juan

For the Iquitos region, the performance of different models is shown in Figure 4. The SVR model has a MAE of 6.09 and an R^2 of 0.1278, showing its adaptability to small samples and nonlinear problems. The LSTM has a MAE of 6.16 and an R^2 of 0.1334, which is better than other deep learning models. The performance of CNN and MLP is relatively poor, with MAEs of 6.29 and 6.58, and R^2

of 0.0222 and 0.1151, respectively. KNN still performs the worst, with a MAE of 6.72 and an R ²of only 0.0317. The results show that for the Iquitos region, the performance of the deep learning model is slightly inferior to that of the SVR, while the performance of the tree model is average.

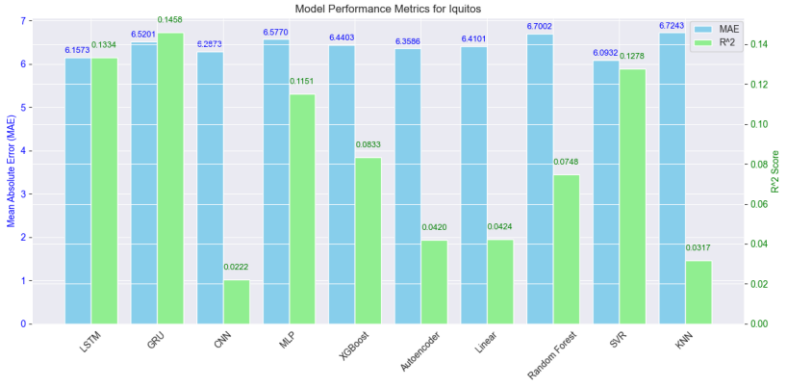

Figure 4. Comparison of MAE and R ²of various models in Iquitos

4.2 Time-series Prediction Results

The time series prediction and actual comparison of San Juan and Iquitos are shown in Figure 5. As can be seen from the figure, in the time series prediction of San Juan, the XGBoost model has a certain ability to capture the peak, but the prediction of extreme values is slightly biased, especially when the number of cases surges, there is an underestimation phenomenon. For Iquitos, the SVR models can track seasonal changes well, but the performance is unstable in the stage with low case numbers, which may be disturbed by noise data.

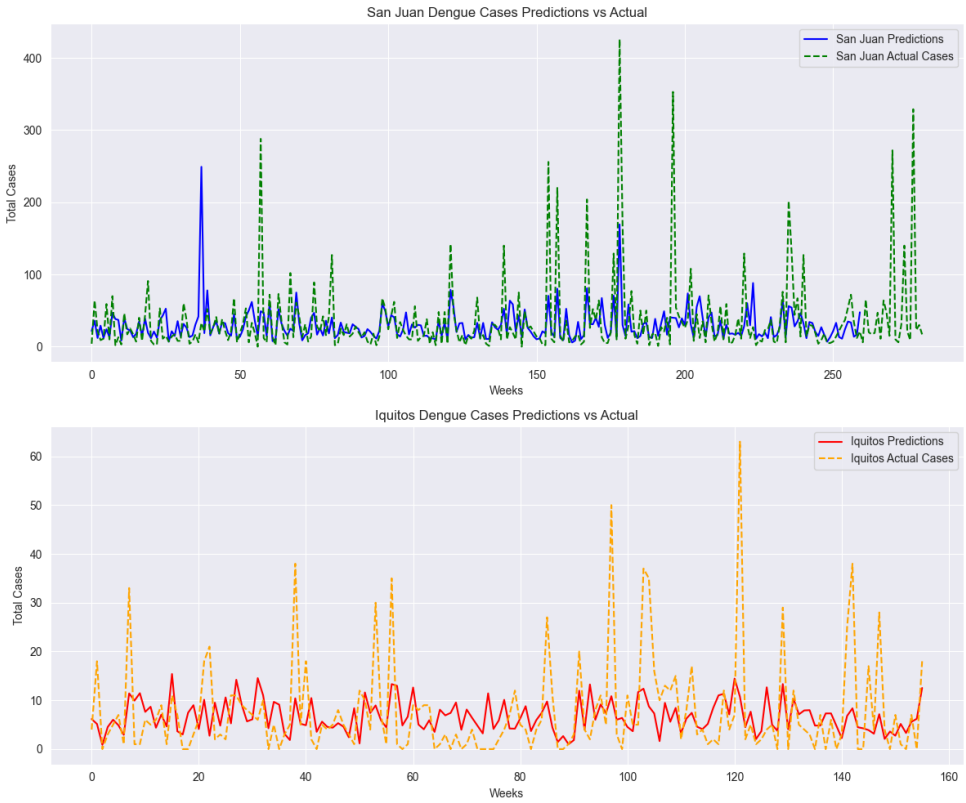

Figure 5. Comparison of time series prediction and actual values in San Juan and Iquitos

4.3 Residual Plot Analysis

In the 3D residual diagram of San Juan (shown in Figure 6), it can be observed that the distribution of the residuals shows certain regularity. In areas with low case numbers, the model's prediction errors are relatively small and evenly distributed. However, as the number of actual cases increases, the volatility of the prediction residuals increases significantly. Especially during the peak period of the epidemic, the model has obvious underestimation, and the residual values show negative values. This shows that although the XGBoost model has a certain ability to fit the overall data in the San Juan region, its prediction performance under extreme values and rapid growth trends still needs to be improved.

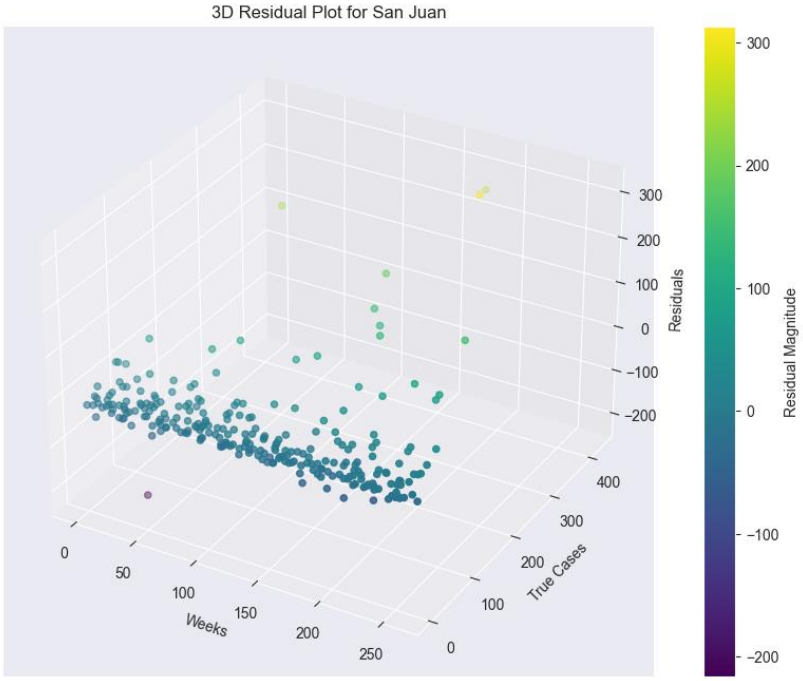

Figure 6. 3D Residual Plot for San Juan

In the 3D residual plot of Iquitos (shown in Figure 7), the distribution characteristics of the residuals are different from San Juan. Overall, the residuals are more scattered in the low-case number stage, which may be caused by the interference of noisy data; while in the medium- and above-case number stage, the fluctuation range of the residuals is reduced, and the model can better capture the Trends in actual cases. However, the residuals of individual extreme points are still large, indicating that the SVR model is not robust enough when dealing with a small number of outliers.

By comparisons, the following conclusions can be drawn: First, the two selected models have certain advantages in capturing the overall trend, but there are obvious shortcomings in dealing with extreme values and rapid changes; secondly, different Differences in regional data characteristics have a significant impact on the prediction performance of the model. For San Juan, the model needs to be further optimized to reduce underestimation during the peak period; for Iquitos, reducing noise interference in the low case number stage will be the key to improving model performance.

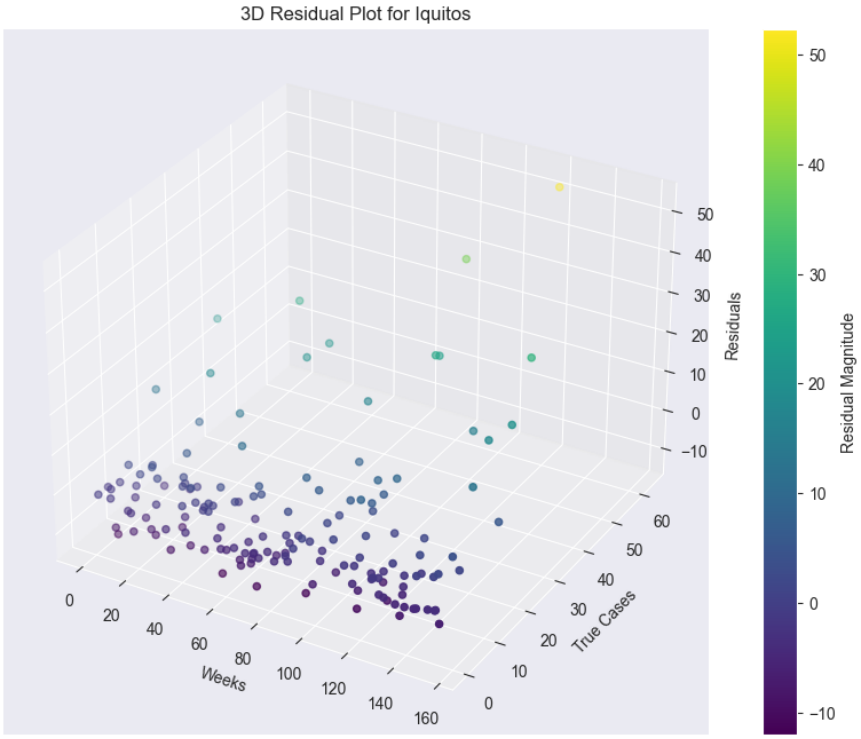

Figure 7. 3D Residual Plot for Iquitos

5. Discussion

Dengue fever cases in the San Juan area show seasonal fluctuation characteristics, and its infection peak is closely related to hot and humid climate conditions. The spread of dengue fever in Iquitos is relatively stable, with weaker seasonal fluctuations in the number of cases and more dependent on the long-term stability of specific environmental variables.

The performance of the models shows obvious regional differences. Among them, the advantages of XGBoost on the San Juan dataset are mainly reflected in its high efficiency in modeling nonlinear relationships and its ability to automatically assign feature importance weights. Although overfitting has been minimized by adjusting hyperparameters such as max_depth and eta, its deviation in extreme value prediction exposes the natural limitations of the gradient boosting algorithm for long-tail distribution responses. This deviation may be due to the fact that the uniform distribution weight allocation of the overall error fails to explicitly distinguish the influence of outliers during the optimization of the objective function. In addition, although the feature subsampling strategy effectively improves the generalization ability, its ability to capture complex interactions between dynamic features is still insufficient.

The relatively low R^2 values observed across all models (maximum 0.31 for San Juan and 0.13 for Iquitos) reflect the inherent complexity and stochasticity of dengue transmission dynamics. Several factors contribute to this limitation. First, dengue transmission involves intricate human-vector-virus interactions that extend beyond climatic variables, including human mobility patterns, vector control interventions, population immunity, and socioeconomic factors not captured in this dataset. Second, weekly case reporting may introduce noise through underreporting and diagnostic delays. Third, the lag between climatic conditions and disease manifestation creates temporal uncertainty. Despite these limitations, the models achieve practical utility for outbreak detection, as evidenced by consistently low MAE values. Similar R^2 ranges (0.15-0.35) have been reported in comparable studies[6][1], suggesting these values are within expected performance bounds for climate-driven dengue

prediction.

In the Iquitos dataset, SVR performs well because its kernelization strategy can better cope with the uncertainty of small sample distribution, especially under the joint optimization of C and epsilon, the fitting of stationary time series reaches the local optimum. However, the limited improvement of its R^2 index exposes the insufficient global fitting ability of support vector regression in high-dimensional feature nesting, especially under the non-balanced feature weight distribution, it is difficult to unify the response to low-dimensional noise and high-frequency oscillation. In addition, the linear growth of SVR in computational complexity makes it show obvious disadvantages in large-scale feature matrices, limiting its scalability.

Deep learning models (such as LSTM and GRU) fail to fully reflect the advantages of time series in San Juan data, indicating that they have deep structural optimization problems in capturing non-stationary dynamic features. Although the regularization strategy based on Dropout and BatchNorm is adopted, the insufficient capture of temporal dependency caused by gradient vanishing during training is not completely solved. In addition, the standardization of feature input (MinMaxScaler) may have a non-negligible interference on the relative weights of high-dimensional features while alleviating the problem of imbalanced input dimensions, weakening the fitting ability of neural networks in specific local areas.

Feature engineering for the San Juan and Iquitos data shows that the generation of lagged features (e.g., 1 week, 2 weeks, 4 weeks) and seasonal feature binarization can significantly improve model performance. However, the fixed setting of the time window length may lead to the loss of high-frequency signals, and the lack of dynamic window adjustment further limits the flexibility of the model at different time scales. In particular, in the Iquitos dataset, due to the weak linear correlation between features, the seasonal feature binarization based on fixed partitions may fail to capture the potential interactions of local complex infectious dynamics.

When predicting extreme values, all models show a clear tendency to underestimate. This underestimation is not only due to the scarcity of extreme value samples in the training set, but also reflects that the existing models fail to fully reflect the weight distribution mechanism of high-risk events when optimizing the objective function. In certain cases, the introduction of objective functions based on quantile regression or uncertainty quantification may have a significant impact on the performance improvement of extreme value prediction. In addition, multimodal data fusion (such as the introduction of socioeconomic and epidemiological data) and the optimization of dynamic feature embedding strategies may significantly enhance the overall stability and local responsiveness of the model.

6. Conclusion

An in-depth exploration of the capabilities of deep learning models, support vector regression, and XGBoost in predicting the number of dengue cases, and a detailed evaluation of their prediction accuracy. The study found that the XGBoost model showed the best performance on the San Juan dataset, while support vector regression was more suitable for the Iquitos dataset. This reflects the difference in adaptability of different models to different data characteristics, and also reminds us The specific circumstances of the data should be fully considered when selecting a model. It is gratifying that feature engineering technology can significantly improve the prediction ability of the model. By introducing techniques such as seasonal binarization and lag features, the seasonal patterns and time dependence of dengue fever incidence are effectively captured, thereby improving the prediction accuracy of the model. However, there are still shortcomings in predicting extreme values, which shows that there is still room for improvement in existing models when dealing with abnormal situations such as epidemic outbreaks.

Future research should be devoted to constructing more sophisticated feature engineering strategies that dynamically adjust the length of the time window to adapt to dynamic changes in the epidemic, and incorporate variables such as socioeconomic demographics to improve the model's ability to capture complex factors. Researchers can also try to explore combining the advantages of different models to build a more powerful and stable prediction system. This study provides a valuable reference for using machine learning technology to predict dengue fever epidemics and lays the foundation for building a more accurate prediction model. These models will help more effectively monitor, predict and control dengue outbreaks, thereby contributing to improving public health.

References

[1] Lowe, R., Bailey, T. C., Stephenson, D. B., Graham, R. J., Coelho, C. A., Carvalho, M. S., & Barcellos, C. (2011). Spatio-temporal modelling of climate-sensitive disease risk: Towards an early warning system for dengue in Brazil. Computers & Geosciences, 37(3), 371-381. https://doi.org/10.1016/j.cageo.2010.01.008

[2] Xu, L., Stige, L. C., Chan, K. S., Zhou, J., Yang, J., Sang, S., ... & Stenseth, N. C. (2017). Climate variation drives dengue dynamics. Proceedings of the National Academy of Sciences, 114(1), 113-118. https://doi.org/10.1073/pnas.1618558114

[3] Hii, Y. L., Rocklöv, J., Ng, N., Tang, C. S., Pang, F. Y., & Sauerborn, R. (2009). Climate variability and increase in intensity and magnitude of dengue incidence in Singapore. Global Health Action, 2(1), 2036. https://doi.org/10.3402/gha.v2i0.2036

[4] Gharbi, M., Quenel, P., Gustave, J., Cassadou, S., La Ruche, G., Girdary, L., & Marrama, L. (2011). Time series analysis of dengue incidence in Guadeloupe, French West Indies: Forecasting models using climate variables as predictors. BMC Infectious Diseases, 11(1), 166. https://doi.org/10.1186/1471-2334-11-166

[5] Leung, X. Y., Islam, R. M., Adhami, M., Ilic, D., McDonald, L., Palawaththa, S., ... & Karim, M. N. (2023). A systematic review of dengue outbreak prediction models: Current scenario and future directions. PLOS Neglected Tropical Diseases, 17(2), e0010631. https://doi.org/10.1371/journal.pntd.0010631

[6] Chen, X., & Moraga, P. (2025). Forecasting dengue across Brazil with LSTM neural networks and SHAP-driven lagged climate and spatial effects. BMC Public Health, 25, 1-22. https://doi.org/10.1186/s12889-025-22106-7

[7] 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. https://doi.org/10.1186/s41182-025-00723-7

[8] Phan, T. H., Nguyen, T. H., Tran, T. T., & Nguyen, H. M. (2024). Leveraging climate data for dengue forecasting in Ba Ria Vung Tau Province, Vietnam: An advanced machine learning approach. Toxics, 9(10), 250. https://doi.org/10.3390/toxics9100250

[9] Bui, H. T. P., Tran, T. N. D., Kubo, T., Iwasaki, C., Pham, D. M., Nguyen, T. T. T., & Yamamoto, T. (2022). Deep learning models for forecasting dengue fever based on climate data in Vietnam. PLOS Neglected Tropical Diseases, 16(6), e0010509. https://doi.org/10.1371/journal.pntd.0010509

[10] Moreira, K. F. A., Oliveira, L. S., Horta, M. A. P., & Magalhães, M. A. F. M. (2024). Forecasting dengue in Bangladesh using meteorological variables with a novel feature selection approach. Scientific Reports, 14, 31234. https://doi.org/10.1038/s41598-024-83770-0