International Journal of Biometeorology (2024) 68:401–410
https://doi.org/10.1007/s00484-023-02605-1

REVIEW PAPER

Meteorological factors cannot be ignored in machine learning-based methods for predicting dengue, a systematic review

Lanlan Fang[1] · Wan Hu[1] · Guixia Pan[1,2]

Received: 30 August 2022 / Revised: 18 September 2023 / Accepted: 13 December 2023 / Published online: 27 December 2023

Abstract

In recent years, there has been a rapid increase in the application of machine learning methods about predicting the incidence of dengue fever. However, the predictive factors and models employed in different studies vary greatly. Hence, we conducted a systematic review to summarize machine learning methods and predictors in previous studies. We searched PubMed, ScienceDirect, and Web of Science databases for articles published up to July 2023. The selected papers included not only the forecast of dengue incidence but also machine learning methods. A total of 23 papers were included in this study. Predictive factors included meteorological factors (22, 95.7%), historical dengue data (14, 60.9%), environmental factors (4, 17.4%), socioeconomic factors (4, 17.4%), vector surveillance data (2, 8.7%), and internet search data (3, 13.0%). Among meteorological factors, temperature (20, 87.0%), rainfall (20, 87.0%), and relative humidity (14, 60.9%) were the most commonly used. We found that Support Vector Machine (SVM) (6, 26.1%), Long Short-Term Memory (LSTM) (5, 21.7%), Random Forest (RF) (4, 17.4%), Least Absolute Shrinkage and Selection Operator (LASSO) (2, 8.7%), ensemble model (2, 8.7%), and other models (4, 17.4%) were identified as the best models based on evaluation metrics used in each article. These results indicate that meteorological factors are important predictors that cannot be ignored and SVM and LSTM algorithms are the most commonly used models in dengue fever prediction with good predictive performance. This review will contribute to the development of more robust early dengue warning systems and promote the application of machine learning methods in predicting climate-related infectious diseases.

Keywords Machine learning · Dengue · Meteorological factors · SVM · LSTM

Introduction

Dengue is an acute infectious disease with Aedes mosquitoes as the main vector (Guzman et al. 2016; Guzman and Harris 2015). The World Health Organization (WHO) reported that dengue was one of the top ten global health threats in 2019. Almost 4 billion people in 128 countries were at risk of dengue (Brady et al. 2012). A study estimated that since 2013, roughly 390 million people have been infected with dengue, of which 96 million had clinical symptoms every year (Bhatt et al. 2013). Since the COVID-19 epidemic, the incidence of dengue fever has risen sharply in many countries, like Singapore. In the past, dengue was mostly confined to tropical and subtropical regions, but now its impact rapidly expanded to the world. Therefore, the establishment of an accurate and early dengue prediction system has been an issue that many scholars have paid more and more attention to.

The machine-learning methods have a non-parametric and non-linear modeling structure (Scavuzzo et al. 2018). They are independent of a priori standards of variable relations and adapted to high-dimensional data, which improve the model's predictive ability with unprecedented accuracy (Bi et al. 2019; DeGregory et al. 2018; Heo et al. 2019). Machine learning methods have been used to successfully predict the incidence of infectious diseases (Kane et al. 2014; Abbasi and Goldenholz 2019; Jiang et al. 2018).

Lanlan Fang and Wan Hu contributed equally to this work.

✉ Guixia Pan
pgxkd@163.com

[1] Present Address: Department of Epidemiology and Biostatistics, School of Public Health, Anhui Medical University, 81 Meishan Road, Hefei 230032, Anhui, China

[2] The Inflammation and Immune Mediated Diseases Laboratory of Anhui Province, Anhui Medical University, Hefei, China

The application of machine learning to predict dengue incidence has also seen a sharp rise. Salim et al. used the support vector machine (SVM) to successfully predict the dengue outbreak based on meteorological factors in Selangor Malaysia (Salim et al. 2021). Zhao et al. developed the random forest (RF) to estimate weekly dengue cases in Colombia for the next 12 weeks using climatic and socioeconomic factors (Zhao et al. 2020). Recently, as a branch of machine learning, deep learning methods such as convolutional neural network (CNN) and long short-term memory (LSTM) have attracted widespread attention in solving various problems (Yousef et al. 2023; Ahmad et al. 2022; Chopra et al. 2022; Khan et al. 2022). Nguyen et al. used CNN and LSTM algorithms to predict dengue based on climate data in Vietnam (Nguyen et al. 2022). Xu et al. proposed an LSTM model with historical dengue cases and climatic factors, which was effective in predicting monthly dengue cases in 20 cities in mainland China (Xu et al. 2020). Although these models were successfully used to predict dengue incidence, the model and the predictors in different studies were not consistent. Hence, we conducted a systematic review to summarize machine learning methods and predictors in previous literature for predicting dengue incidence based on machine learning methods.

The contribution of this review:

This systematic review aims to review all the published literature on machine learning models for dengue fever prediction. It provides a basis for developing more robust dengue early warning systems in the future by summarizing the characteristics of predictor variables and prediction methods.

Methods

Search strategy

This systematic review was conducted in compliance with the recommendations of the PRISMA guidelines. We searched PubMed, ScienceDirect, and Web of Science databases for articles published up to July 2023. ((Machine learning OR deep learning OR Ensemble OR RF OR SVM OR LASSO OR LSTM) AND (dengue)) was used as the search term.

Eligibility criteria

The eligibility criteria were employed to select suitable papers in the search results in order to make the selected articles as complete and uniform as possible. Three inclusion criteria were as follows: first, this review included only peer-reviewed journal articles published in English full text/pdf, making the information gained authoritative. Second, papers had to incorporate the forecast of dengue. Third, papers must include the machine learning method. Three exclusion criteria were as follows: first of all, abstract-only reports and duplicate titles were omitted. Moreover, papers containing unspecified or suspected dengue were excluded. In addition, papers reporting dengue infection in nonhuman cases were also removed.

Outcome

The outcomes were the incidence rate or cases of dengue as well as dengue serotype (e.g., dengue fever (DF) and dengue hemorrhagic fever (DHF)). The outcomes were converted into the categorical variable according to whether the incidence rate or cases of dengue exceeded a certain threshold when evaluating whether an outbreak of dengue occurred.

Data extraction and analysis

The relevant data were extracted by two independent investigators (Lanlan Fang and Wan Hu): author, year of publication, study area, study period, outcome, predictors, methods used, model validation, evaluation index, and optimal model. Any disagreement was resolved by discussion and consensus with a senior reviewer (Guixia Pan) if needed.

The above key information was extracted from the included studies to make a table. Descriptive statistics were performed based on the characteristics of the predictors and models (including model validation techniques and evaluation metrics).

Results

Literature search

We found 4,517 records in the search and deleted 932 duplicates. The initial search result was 3,585 articles, of which 1022 were considered potentially relevant and required further reading of the abstract for screening. After reading the full text of 126 articles, it was found that 23 of them fully met the eligibility criteria. The screening process is shown in Fig. 1. Basic information of 23 studies including author (year), study area (period), outcome, predictors, method used, model validation, evaluation index, and optimal model are presented in Table 1.

Study characteristics

Figure 2 summarizes the types of predictors used in the 23 included articles according to Table 1. 95.7% of the studies utilized meteorological factors. 60.9% of the studies incorporated historical dengue data. 17.4% of the studies considered

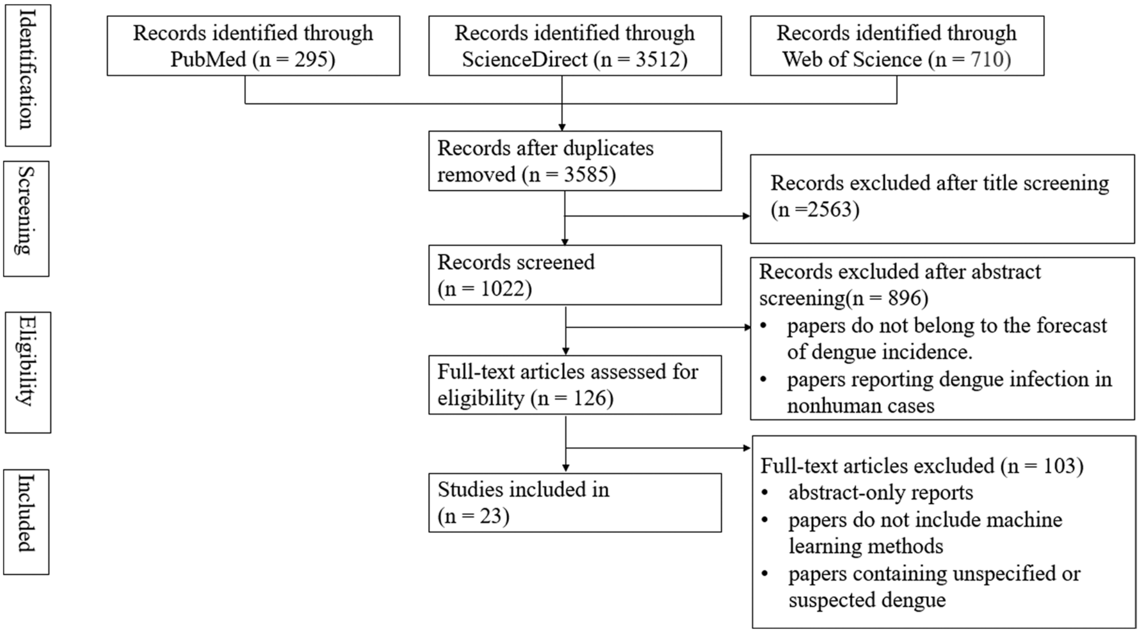
Identification
Screening
Eligibility
Included
Records identified through PubMed (n = 295)
Records identified through ScienceDirect (n = 3512)
Records identified through Web of Science (n = 710)
Records after duplicates removed (n = 3585)
Records excluded after title screening (n =2563)
Records screened (n = 1022)
Records excluded after abstract screening(n = 896)
• papers do not belong to the forecast of dengue incidence.
• papers reporting dengue infection in nonhuman cases
Full-text articles assessed for eligibility (n = 126)
Full-text articles excluded (n = 103)
• abstract-only reports
• papers do not include machine learning methods
• papers containing unspecified or suspected dengue
Studies included in (n = 23)

Fig. 1 Flowchart of literature search

environmental factors. 17.4% of the studies took into account socio-economic factors. 8.7% of the studies utilized vector surveillance data. 13.0% of the studies considered internet search data. Figure 3 summarizes the combination of predictors used in the 23 included articles. 7 articles (30.4%) were based exclusively on meteorological predictors. 15 articles (65.2%) used a combination of meteorological and other types of factors. Of these studies, 21.7% used meteorological factors and historical dengue cases. 8.7% used a combination of meteorological factors, historical dengue cases, and environmental factors. 8.7% considered meteorological factors, historical dengue cases, and socioeconomic factors. 8.7% incorporated meteorological factors, historical dengue cases, and internet search data. 4.3% considered meteorological and environmental factors. 4.3% used meteorological, environmental, and socioeconomic factors. 8.7% integrated meteorological factors, historical dengue cases, vector monitoring data, and socioeconomic factors. Only one article (4.3%) did not involve meteorological factors, which was based on historical dengue cases and internet search data.

Table 2 lists the specific predictors that appeared in the prediction models of the 23 studies. Among meteorological factors, the majority of studies utilized temperature (20, 87.0%), rainfall (20, 87.0%), and relative humidity (14, 60.9%). Some studies also included wind speed (4, 17.4%), atmospheric pressure (2, 8.7%), evaporation (2, 8.7%), and diurnal temperature range (DTR) (1, 4.3%). In terms of historical dengue data, 43.5% of the studies utilized weekly data, 13.0% used monthly data, and 8.7% used yearly data. Environmental factors included enhanced vegetation index (EVI) (2, 8.7%), normalized difference vegetation index (NDVI) (2, 8.7%), southern oscillation index (SOI) (1, 4.3%), and flood occurrence (1, 4.3%). Socio-economic factors consisted of the population density (3, 13.0%), Gini index (1, 4.3%), education coverage rate (1, 4.3%), and monthly garbage collection quantity (1, 4.3%). Vector surveillance data included the Aedes aegypti index (1, 4.3%) and the Breteau index (1, 4.3%). In addition, internet search data comprised Tweets (1, 4.3%), Baidu search queries (1, 4.3%), and Google search queries (1, 4.3%).

Figure 4A summarizes the techniques of model validation. 37.5% employed split-sample validation, 29.2% used cross-validation, 25.0% utilized out-of-sample validation, and 8.3% employed retrospective validation. Figure 4B summarizes the model evaluation metrics. These included mean absolute error (MAE) (21.2%), root mean square error (RMSE) (18.2%), accuracy (18.2%), sensitivity (6.1%), specificity (6.1%), area under curve (AUC) (9.1%), mean absolute percentage error (MAPE) (3.0%), mean percentage error (MPE) (3.0%), Theil's coefficient (3.0%), Matthew Correlation Coefficient (MCC) (3.0%), R-squared (3.0%), mean squared error (MSE) (3.0%), and mean (3.0%). Figure 4C summarizes the optimal models judged based on evaluation metrics in each article. The best models include SVM (6, 26.1%), LSTM (5, 21.7%), RF (4, 17.4%), LASSO (2, 8.7%), ensemble model (2, 8.7%), and other machine learning methods (4, 17.4%).

Table 1 The basic information about 23 studies involved in the review of machine-learning-based prediction models for dengue incidence

References	Study area (Period)	Outcome	Predictors	Methods used	Model validation	Evaluation index	Optimal model
(Althouse et al. 2011)	Singapore (2004–2011) and Bangkok (2004–2011)	weekly and monthly dengue incidence	Historical dengue cases and search query data.	SVM and statistical model	multiple cross-validations	AUC	SVM
(Kesorn et al. 2015)	Three provinces in central Thailand (2007-2017)	monthly dengue incidence	Meteorological factors, historical dengue cases, vector surveillance data, and socioeconomic factors.	the SVM with different kernels, the KNN, DT, and the NN	10-fold cross-validation	Accuracy	SVM (RBF)
(Stolerman et al. 2019)	seven Brazilian state capitals (2002–2017)	yearly dengue incidence	Meteorological factors	SVM-R and SVM-L	Split sample validation	Accuracy	SVM (RBF)
(Salim et al. 2021)	five districts in Selangor, Malaysia (2013-2017)	weekly dengue cases	Meteorological factors	SVM with different kernels, DT, ANN, and GNB	Split sample validation	Accuracy, Sensitivity, Specificity	SVM (linear kernel)
(Mustaffa and Yusof 2011)	five districts in Selangor, Malaysia (2004-2005)	weekly dengue cases	Meteorological factors and historical dengue cases.	SVM and ANN	out-of-sample validation	MSE, Accuracy	SVM
(Guo et al. 2017)	Guangdong (2011-2014)	weekly dengue cases	Meteorological factors, historical dengue cases, and search query data.	SVM, GBM, LASSO, and statistical models	cross-validation	R-squared	SVM
(Mussumeci and Codeco Coelho 2020)	790 Brazilian cities (2010-2018)	weekly dengue incidence	Meteorological factors, historical dengue cases, and search query data.	LSTM, RF, and LASSO	Retrospective validation and out-of-sample validation	MPE	LSTM
(Xu et al. 2020)	20 Chinese Cities (2005-2018)	monthly dengue cases	Meteorological factors and historical dengue cases	LSTM, LSTM-TL, BPNN, SVM, GBM, and statistical models	Split sample validation	RMSE	LSTM-TL
(Li et al. 2022)	Brazil (2007-2019)	weekly dengue cases	Meteorological factors, environment, and historical dengue cases	LSTM	Split sample validation	RMSE, MAE	LSTM
(Li 2022)	Brazil (2013-2020)	weekly dengue cases	Meteorological factors, environment, and historical dengue cases	RF, LSTM, and LSTM-ATT	Retrospective validation	RMSE, MAE	LSTM-ATT
(Nguyen et al. 2022)	Vietnam (1997 to 2013)	yearly dengue case	Meteorological factors	CNN, LSTM, LSTM-ATT	Split sample validation (2014-2016)	MAE, RMSE	LSTM-ATT
(Shi et al. 2016)	Singapore (2001-2010)	weekly dengue incidence	Meteorological factors, historical dengue cases, vector surveillance data, and socioeconomic factors.	LASSO and statistical models	out-of-sample validation	MAPE	LASSO
(Chen et al. 2018)	Singapore (2010-2016)	weekly dengue cases	Meteorological factors and historical dengue cases	LASSO	Split sample validation	AUC	LASSO
(Benedum et al. 2020)	Iquitos, San Juan, and Singapore; (1990-2016)	weekly dengue cases	Meteorological factors, historical dengue cases and socioeconomic factors.	RF, and statistical models	Split sample validation	MAE, MCC	RF

Table 1 (continued)

References	Study area (Period)	Outcome	Predictors	Methods used	Model validation	Evaluation index	Optimal model
(Zhao et al. 2020)	Colombia (2014-2018)	weekly dengue cases	Meteorological factors, environment, and socio-economic factors.	RF and ANN	leave-one-season-out cross-validation	MAE	RF
(Carvajal et al. 2018)	Metropolitan Manila (2009-2017)	weekly dengue incidence	Meteorological factors and environment	RF, GBM, and statistical model	out-of-sample validation	MAE	RF
(Gupta et al. 2023)	San Juan (1990-2008), Iquitos (2000-2010)	weekly dengue cases	Meteorological factors	RF, DT, GNB, SVM, KNN	10-fold cross-validation	Mean	RF
(Buczak et al. 2018)	Iquitos (2000-2009) and San Juan (1990-2009)	weekly dengue case	Meteorological factors and historical dengue cases	Ensemble approach	Split sample validation (2009-2013)	RMSE, MAE	Ensemble approach
(McGough et al. 2021)	Brazil (2001-2017)	yearly dengue incidence	Meteorological factors and historical dengue cases	Ensemble approach	out-of-sample validation	Accuracy, Sensitivity, Specificity	Ensemble approach
(Lowe et al. 2018)	in Barbados (2010–2017)	monthly dengue incidence	Meteorological factors	Other	Leave-one-season-out cross-validation	AUC	Other
(Bett et al. 2019)	in Vietnam (2001-2012)	monthly dengue incidence	Meteorological factors	Other	Split sample validation	Theil’s coefficient	Other
(Anno et al. 2019)	in southwest Taiwan (1998-2015)	monthly dengue cases	Meteorological factors	Other	8-fold cross-validation	Accuracy	Other
(Jain et al. 2019)	the fifty districts of Thailand (2017-2015)	monthly dengue cases	Meteorological factors, historical dengue cases, and socioeconomic factors.	Other	out-of-sample validation	RMSE	Other

Abbreviations, *ANN,* artificial neural networks; *MAE*, mean absolute error; *SVR*, support vector regression; *SVM-L,* SVM-Linear; *SVM-P,* SVM-Polynomial; *SVM-R*, SVM-RBF; *RF,* random forest; *GBM,* gradient boosting machine; *GNB*, gaussian Naive Bayes; *AUC,* area under the ROC curve; *ROC*, relative (receiver) operating characteristic; *MCC,* Matthew’s correlation coefficient; *NN*, neural networks; *KNN,* K-nearest neighbor; *CNN,* convolutional neural network; *BPNN,* back propagation neural network; *DT,* decision tree; *MPE,* mean percentage error; *MSE,* mean squared error; *MAPE,* mean absolute percentage error; *RMSE,* root mean square error; *TL,* transfer learning; *LSTM,* long short-term memory; *LSTM-ATT,* attention-enhanced LSTM; and *LASSO*, least absolute shrinkage and selection operator

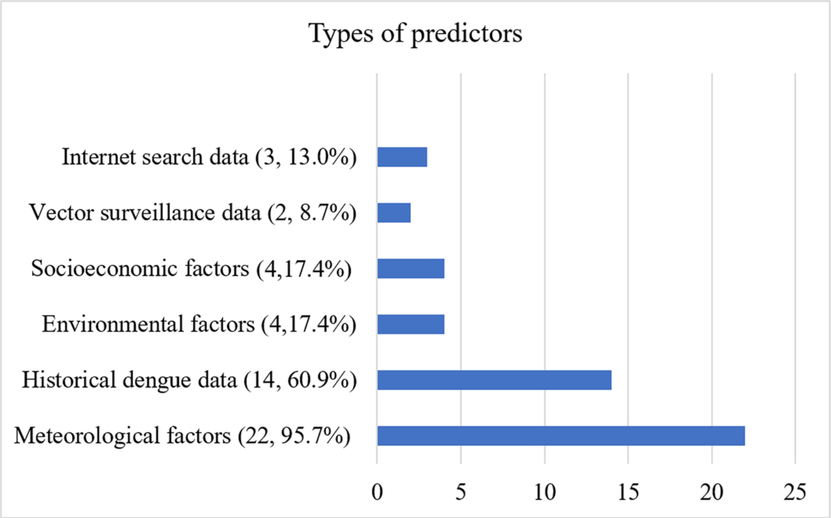

Fig. 2 The types of predictors used in the 23 included articles. * Since multiple predictors may be used in a single article, the sum of the factors did not equal 23 and the sum of the percentages did not equal 100

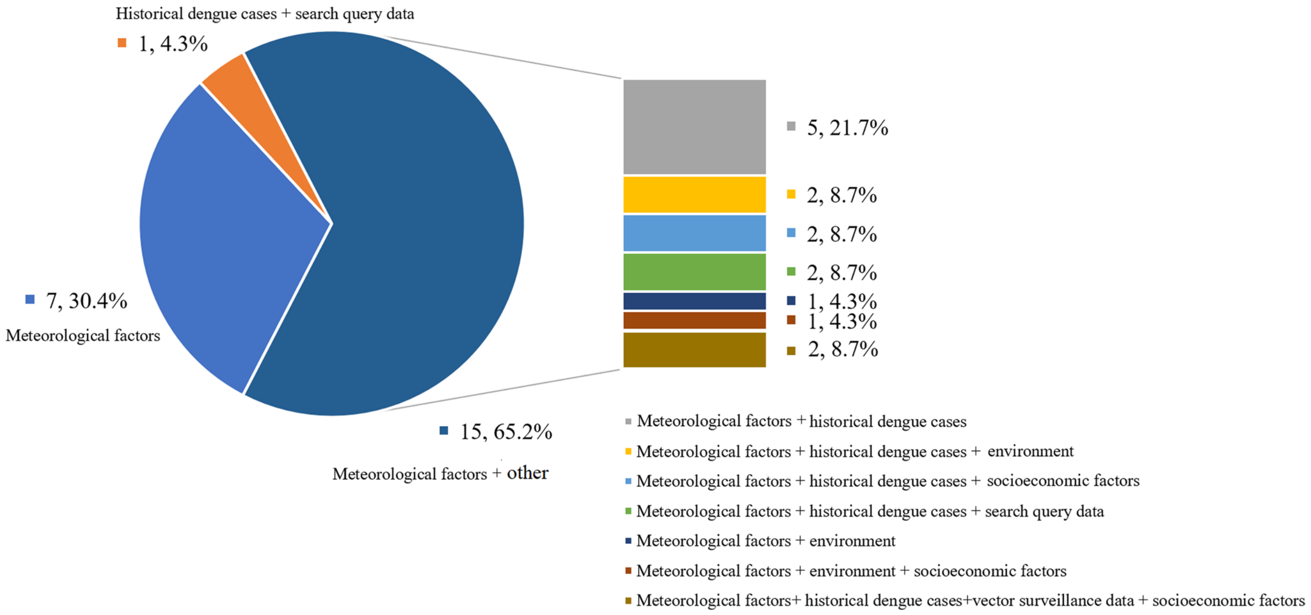

Fig. 3 The combinations of predictors used in the 23 included articles

Literature review

We respectively presented the application of SVM, LSTM, RF, LASSO, and ensemble models in the context of dengue epidemiology.

The SVM can solve classification and regression prediction (Huang et al. 2018, Noble 2006). A total of 6/23 articles identified SVM as the best model, of which 4 were for classification and 2 for regression. For classification prediction, Althouse et al. defined the binary outcome as 1 for high-incidence periods and 0 for other periods. The AUC of the SVM model was 0.906 in Singapore and 0.960 in Bangkok (Althouse et al. 2011). Kesorn et al. developed an SVM with an RBF kernel to predict high-prevalence areas in central Thailand with the highest accuracy (96.26%) compared to other models (Kesorn et al. 2015). When Stolerman et al. identified the occurrence of dengue epidemic years, the accuracy of SVM with RBF kernel (91%) was higher than the accuracy of SVM with linear kernel (82%) (Stolerman et al. 2019). Salim et al. used several machine learning methods to assess whether dengue outbreaks would occur, with SVM (linear kernel) providing the best predictions (Salim et al. 2021). For regression prediction, dengue case counts or incidence is a continuous dependent variable. Mustaffa developed an SVM model with the RBF kernel to predict the dengue case counts.

Table 2 The predictors that appeared in the prediction models of the 23 studies (number, percentage)

Meteorological factors	Historical dengue data	Environmental factors	Socioeconomic factors	Vector surveillance data	Internet search data
Temperature (20, 87.0%)	Weekly cases (10, 43.5%)	EVI (2, 8.7%)	Population (3, 13.0%)	Ades aegypti index (1, 4.3%)	Tweets (1, 4.3%)
Rainfall (20, 87.0%)	Monthly cases (3, 13.0%)	NDVI (2, 8.7%)	Gini index (1, 4.3%)	Breteau index (1, 4.3%)	Baidu search queries (1, 4.3%)
Humidity (14, 60.9%)	Yearly cases (2, 8.7%)	SOI (1, 4.3%)	Education coverage (1, 4.3%)		Google search queries (1, 4.3%)
Wind speed (4, 17.4%)		Flood occurrence (1, 4.3%)	Monthly garbage collection (1, 4.3%)		
Pressure (2, 8.7%)					
Evaporation (2, 8.7%)					
DTR (1, 4.3%)					

EVI, Enhanced vegetation index; *SOI,* Southern oscillation index; *NDVI,* Normalized difference vegetation index; *DTR,* Diurnal temperature range

*Since multiple predictors may be used in a single article, the sum of the factors does not equal 23 and the sum of the percentages does not equal 100

The MSE and accuracy of this model were 0.0063 and 86.84% (Mustaffa). Guo et al implemented a support vector regression (SVR) with a linear kernel function to predict the weekly number of dengue cases. (Guo et al. 2017). The R-squared of this model was 0.99, and it achieved the best performance compared to other forecasting techniques.

Recently, deep learning methods have also been increasing in dengue prediction. CNN is widely used to process image data (Salehi et al. 2023) and LSTM is applied to deal with time series problems (Houdt et al. 2020). A total of 5/23 papers had the best performance of the LSTM algorithm compared to other ML tools. Mussumeci and Codeco fitted the models (LSTM, RF, and LASSO) to forecast the weekly dengue incidence, and the LSTM model was the best with an MPE of 0.04 (Mussumeci and Codeco Coelho 2020). Xu et al. developed an LSTM-TL model to effectively predict monthly dengue cases in 20 cities in mainland China (Xu et al. 2020). This model (RMSE=0.91) outperformed other algorithms (LSTM, BPNN, SVM, and GBM). Li et al. fitted an LSTM model for forecasting the number of dengue cases using rainfall, temperature, relative humidity, mean normalized difference vegetation index (NDVI), and historical dengue cases in Brazil (Li et al. 2022). Li further improved this algorithm and constructed the LSTM-ATT model, which has higher prediction performance than RF and LSTM (Li 2022). Nguyen et al. used the deep learning models (CNN, LSTM, LSTM-ATT) for forecasting dengue fever in Vietnam (Nguyen et al. 2022), which also revealed that LSTM-ATT outperformed the base LSTM and CNN.

A total of 4/23 papers had the best performance of the RF compared to other models. Benedum et al. developed an RF algorithm with the best performance to predict dengue case counts (regression) and outbreaks (classification) in 4 to 12 weeks (Benedum et al. 2020). Zhao et al. developed national and departmental RF models and an ANN model to estimate weekly dengue cases for the next 12 weeks in Colombia.

The RF model trained on national data (MAE=24.56) was better than the RF model trained on departmental data (MAE=26.76) and ANN (MAE=25.25) (Zhao et al. 2020). Carvajal et al predicted dengue incidence in Manila based on weather factors and their corresponding lagged effect. The RF model with delayed meteorological effects (MAE=0.15) showed the best predictive accuracy compared to GBM and other statistical models (Carvajal et al. 2018). Gupta et al predicted the dengue case counts in San Juan and Iquitos via various machine learning algorithms. RF produced better results than DT, KNN, SVR, and GNB (Gupta et al. 2023).

A total of 2/23 papers had the best performance of the LASSO algorithm compared to other models. Shi et al. developed a three-month real-time dengue forecast system with the LASSO-derived model. This model forecasted the weekly incidence of dengue and utilized multiple data streams that were updated weekly, including historical dengue case data, climate data, vector surveillance data, and temporal data. The MAPE was the smallest compared with SARIMA, and step-down linear regression (Shi et al. 2016). Chen et al. developed a novel framework based on LASSO regression for producing a spatiotemporal dengue forecast at a neighborhood-level spatial resolution to distinguish between high and low-risk areas (Chen et al. 2018).

The ensemble model combines different models to improve their performance and robustness (Haq et al. 2022). A total of 2/23 papers with the ensemble model were performed. Buczak et al. proposed an ensemble model created by combining three disparate types of component

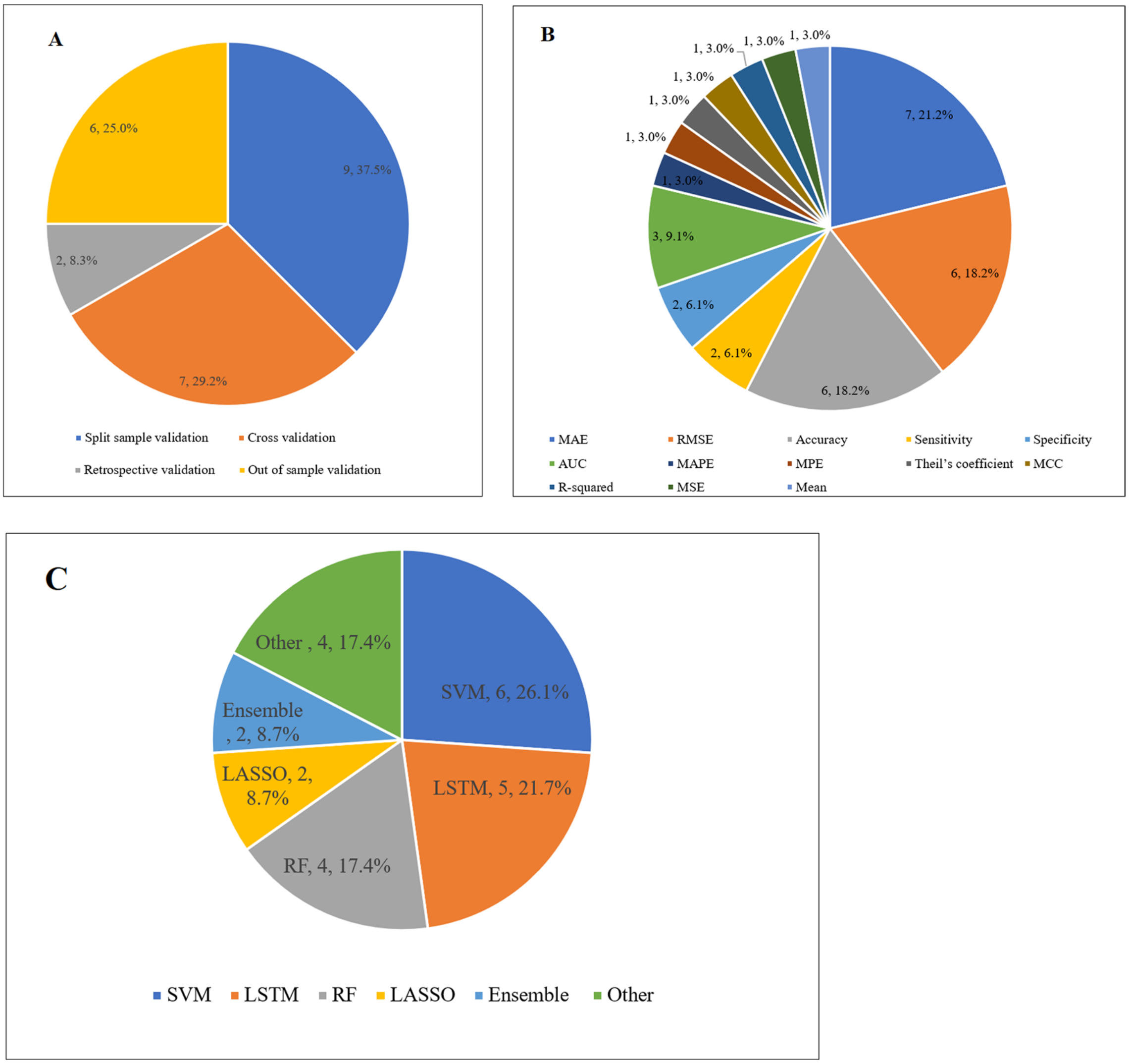

Fig. 4 Model validation techniques (**A**) and model evaluation metrics (**B**) and the optimal models judged based on evaluation metrics in each article (**C**). * Since multiple techniques and evaluation metrics may be used in a single article, the sum of techniques and evaluation metrics did not equal 23

models, which was the best in predicting the total dengue case counts and peak height for Iquitos, Peru (Buczak et al. 2018). McGough et al. developed a high-accuracy ensemble model for predicting the year of dengue outbreaks in Brazil using meteorological factors and historical dengue cases (McGough et al. 2021).

Discussion

Here, we conducted a systematic review to summarize the forecast of dengue incidence based on machine learning methods. We found that 95.65% of the studies incorporated meteorological factors, with the majority utilizing temperature, rainfall, and relative humidity in their machine-learning prediction models. Among the 23 articles, the models most frequently selected as optimal were SVM with 6 articles (26.1%) and LSTM with 5 articles (21.7%).

Machine learning methods are data-driven forecasting methods. In the selection of independent variables, as long as a certain variable can improve the prediction accuracy of the model, it can be introduced into the model regardless of whether it is a risk factor for dengue incidence (Jamshidi et al. 2019; Heo et al. 2019; Peiffer-Smadja et al. 2020). In this review, the predictive variables included not only traditional risk factors that affect the infection of dengue but also other factors that help capture the trend of dengue incidence (i.e., historical dengue data and internet search data (Spratt et al. 2013). First, 95.65% of the articles included meteorological factors, which indicated that meteorological factors were not negligible in predicting the incidence of dengue. Common meteorological indicators included temperature, rainfall, and relative humidity (Wu et al. 2018b, Xiang et al. 2017). In addition, environmental factors (i.e., EVI and NDVI were also considered in some studies (Li et al. 2022) (Zhao et al. 2020), Second, 60.9% of the articles included historical dengue cases. As dengue is an infectious disease, historical cases had a significant impact on the current dengue epidemic. Thus, historical dengue data should be included as a predictor (Ramadona et al. 2016; Jain et al. 2019). Third, 17.4% of studies have also taken into account socioeconomic factors such as population and the Gini index. Fourth, to improve the accuracy of prediction as much as possible, internet search data (i.e., Google Trends, Baidu Index, and Twitter), as currently an emerging predictor, have proven to be a good complement to traditional surveillance data (Wu et al. 2018a). In general, although a wide variety of predictors were incorporated into machine learning models, meteorological factors were non-negligible in the prediction model.

In our review, the SVM was the most frequent (6, 26.1%) optimal model in the 23 included studies. The SVM algorithm has the capability of handling small sample sets, controlling overfitting, and dealing with nonlinear relationships, making it suitable for dengue incidence prediction. Looking at the performance metrics of the developed SVM models, Kesorn et al. obtained the highest accuracy of 96.26 with SVM (RBF) (Kesorn et al. 2015). Althouse et al. reported that the AUC of the SVM model was the highest at 0.960 in Bangkok (Althouse et al. 2011). With the advancement of deep learning methods, LSTM is also becoming increasingly popular, with a total of 5/23 articles being the optimal model. The LSTM has the unique advantage of being able to capture long-term dependencies, predict the incidence of dengue fever at different time intervals, remember important patterns and trends in historical dengue data, and learn complex patterns (nonlinearity and interactions) (Sagheer and Kotb 2019). The LSTM-ATT developed by Nguyen et al. performed the best among multiple deep learning techniques, with an average ranking of 1.60 for RMSE, and 1.95 for MAE (Nguyen et al. 2022). In addition, other algorithms have some unique advantages. The RF can rank the importance of predictors, and Carvajal et al. found that relative humidity, rainfall, and temperature were the best predictors (Carvajal et al. 2018). The LASSO can screen for valuable factors to solve serious collinearity problems, thereby improving the accuracy of predictions (Ranstam and Cook 2018) (Wang et al. 2018). The ensemble model combined the predictions of multiple basic models to obtain a more accurate and robust final prediction. Therefore, in actual prediction, we should not only rely on one method but use multiple methods to fit multiple models and select the model with the best prediction effect according to evaluation criteria.

This review had several limitations. Firstly, our review only included peer-reviewed literature from selected databases and some dengue outbreaks may not have been recorded; therefore, our results should be interpreted with caution. Secondly, most cases of dengue were asymptomatic; thus, there may be some patients with dengue that have not been confirmed and the actual number of cases may be much higher than recorded. Thirdly, because the model evaluation indicators of 17 papers were not the same and different regions varied largely, we cannot use a unified quantitative indicator to determine which machine learning method was the best. Fourthly, in the prediction of dengue incidence, we only discussed machine learning methods. While we did not pay attention to some improved time-series models proven to have higher accuracy.

Conclusions

We reviewed the 23 articles where machine learning methods were successfully applied to predict dengue incidence and confirmed that machine learning methods were attractive enough. The SVM was the most frequent model for dengue prediction with good predictive performance. With the development of deep learning, LSTM might be a more promising model for dengue prediction. More importantly, the meteorological factors including temperature, rainfall, and relative humidity could not be ignored in the prediction model.

Acknowledgments Thank you very much Hu Wan for helping us in the data extraction process.

Authors' contributions FL and PG conceived the study; FL designed the study protocol; FL and WH carried out the literature search and review; FL drafted the manuscript; PG and WH critically revised the manuscript for intellectual content. All authors read and approved the final manuscript. FL, WH, and PG are guarantors of the paper.

Funding This work was supported by the Special Funding for the Prevention and Control of New Coronary Pneumonia Epidemic from the China Postdoctoral Science Foundation (2020T170001ZX).

Data Availability All literature reviewed in this study is openly available and cited in the text.

Declarations

Ethical approval and consent to participate Not applicable.

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

References

Abbasi B, Goldenholz DM (2019) Epilepsia 60:2037–2047

Ahmad S, Khan S, AlAjmi M-F, Dutta A-K, Dang L-M, Joshi G-P, Moon H (2022) Comput Mater Continua 73:965–979

Althouse BM, Ng YY, Cummings DA (2011) PLoS Negl Trop Dis 5:e1258

Anno S, Hara T, Kai H, Lee MA, Chang Y, Oyoshi K, Mizukami Y Tadono T (2019) Geospat Health 14

Benedum CM, Shea KM, Jenkins HE, Kim LY, Markuzon N (2020) PLoS Negl Trop Dis 14:e0008710

Bett B, Grace D, Lee HS, Lindahl J, Nguyen-Viet H, Phuc PD, Quyen NH, Tu TA, Phu TD, Tan DQ, Nam VS (2019) PLoS One 14:e0224353

Bhatt S, Gething PW, Brady OJ, Messina JP, Farlow AW, Moyes CL, Drake JM, Brownstein JS, Hoen AG, Sankoh O, Myers MF, George DB, Jaenisch T, Wint GR, Simmons CP, Scott TW, Farrar JJ, Hay SI (2013) Nature 496:504–507

Bi Q, Goodman KE, Kaminsky J, Lessler J (2019) Am J Epidemiol 188:2222–2239

Brady OJ, Gething PW, Bhatt S, Messina JP, Brownstein JS, Hoen AG, Moyes CL, Farlow AW, Scott TW, Hay SI (2012) PLoS Negl Trop Dis 6:e1760

Buczak AL, Baugher B, Moniz LJ, Bagley T, Babin SM, Guven E (2018) PLoS One 13:e0189988

Carvajal TM, Viacrusis KM, Hernandez LFT, Ho HT, Amalin DM, Watanabe K (2018) BMC Infect Dis 18:183

Chen Y, Ong JHY, Rajarethinam J, Yap G, Ng LC, Cook AR (2018) BMC Med 16:129

Chopra P, Junath N, Singh SK, Khan S, Sugumar R, Bhowmick M (2022) BioMed Res Int 2022:6336700

DeGregory KW, Kuiper P, DeSilvio T, Pleuss JD, Miller R, Roginski JW, Fisher CB, Harness D, Viswanath S, Heymsfield SB, Dungan I, Thomas DM (2018) Obes Rev 19:668–685

Guo P, Liu T, Zhang Q, Wang L, Xiao J, Zhang Q, Luo G, Li Z, He J, Zhang Y, Ma W (2017) PLoS Negl Trop Dis 11:e0005973

Gupta G, Khan S, Guleria V, Almjally A, Alabduallah BI, Siddiqui T, Albahlal BM, Alajlan SA, Al-Subaie M (2023) Diagnostics (Basel) 13

Guzman MG, Harris E (2015) Lancet 385:453–465

Guzman MG, Gubler DJ, Izquierdo A, Martinez E, Halstead SB (2016) Nat Rev Dis Primers 2:16055

Haq AU, Li JP, Agbley BLY, Khan A, Khan I, Uddin MI, Khan S (2022) IEEE J Biomed Health Inform 26:5004–5012

Heo J, Yoon JG, Park H, Kim YD, Nam HS, Heo JH (2019) Stroke 50:1263–1265

Houdt GV, Mosquera C, Nápoles G (2020) Artificial Intell Rev, 5929-5955

Huang S, Cai N, Pacheco PP, Narrandes S, Wang Y, Xu W (2018) Cancer Genomics Proteomics 15:41–51

Jain R, Sontisirikit S, Iamsirithaworn S, Prendinger H (2019) BMC Infect Dis 19:272

Jamshidi A, Pelletier JP, Martel-Pelletier J (2019) Nat Rev Rheumatol 15:49–60

Jiang D, Hao M, Ding F, Fu J, Li M (2018) Acta Tropica 185:391–399

Kane MJ, Price N, Scotch M, Rabinowitz P (2014) BMC Bioinformatics 15:276

Kesorn K, Ongruk P, Chompoosri J, Phumee A, Thavara U, Tawatsin A, Siriyasatien P (2015) PLoS One 10:e0125049

Khan S, Fazil M, Sejwal VK, Alshara MA, Alotaibi RM, Kamal A, Baig AR (2022) J King Saud Univ - Comput Inform Sci 34:4335–4344

Li Z (2022) Int J Environ Res Public Health 19

Li Z, Gurgel H, Xu L, Yang L, Dong J (2022) Biology (Basel) 11

Lowe R, Gasparrini A, Van Meerbeeck CJ, Lippi CA, Mahon R, Trotman AR, Rollock L, Hinds AQJ, Ryan SJ, Stewart-Ibarra AM (2018) PLoS Med 15:e1002613

McGough SF, Clemente L, Kutz JN, Santillana M (2021) J R Soc Interface 18:20201006

Mussumeci E, Codeco Coelho F (2020) Spat Spatiotemporal Epidemiol 35:100372

Mustaffa Z, Yusof Y (2011) Int J Comput Theory Eng 489–493

Nguyen VH, Tuyet-Hanh TT, Mulhall J, Minh HV, Duong TQ, Chien NV, Nhung NTT, Lan VH, Minh HB, Cuong D, Bich NN, Quyen NH, Linh TNQ, Tho NT, Nghia ND, Anh LVQ, Phan DTM, Hung NQV, Son MT (2022) PLoS Negl Trop Dis 16:e0010509

Noble WS (2006) Nat Biotechnol 24:1565–1567

Peiffer-Smadja N, Rawson TM, Ahmad R, Buchard A, Georgiou P, Lescure FX, Birgand G, Holmes AH (2020) Clin Microbiol Infect 26:584–595

Ramadona AL, Lazuardi L, Hii YL, Holmner Å, Kusnanto H, Rocklöv J (2016) PLoS One 11:e0152688

Ranstam JL, Cook JA (2018) British J Surg, 1348

Sagheer A, Kotb M (2019) Neurocomputing, 203-213

Salehi AW, Khan S, Gupta G, Alabduallah BI, Almjally A, Alsolai H, Siddiqui T, Mellit A (2023) Sustainability 15:5930

Salim NAM, Wah YB, Reeves C, Smith M, Yaacob WFW, Mudin RN, Dapari R, Sapri N, Haque U (2021) Sci Rep 11:939

Scavuzzo JM, Trucco F, Espinosa M, Tauro CB, Abril M, Scavuzzo CM, Frery AC (2018) Acta Tropica 185:167–175

Shi Y, Liu X, Kok SY, Rajarethinam J, Liang S, Yap G, Chong CS, Lee KS, Tan SS, Chin CK, Lo A, Kong W, Ng LC, Cook AR (2016) Environ Health Perspect 124:1369–1375

Spratt H, Ju H, Brasier AR (2013) Methods 61:73–85

Stolerman LM, Maia PD, Kutz JN (2019) PLoS One 14:e0220106

Wang S, Ji B, Zhao J, Liu W, Xu T (2018) Transportation Research: Part D:817-824

Wu C, Kao SC, Shih CH, Kan MH (2018a) Acta Trop 183:1–7

Wu X, Lang L, Ma W, Song T, Kang M, He J, Zhang Y, Lu L, Lin H, Ling L (2018b) Sci Total Environ 628-629:766–771

Xiang J, Hansen A, Liu Q, Liu X, Tong MX, Sun Y, Cameron S, Hanson-Easey S, Han GS, Williams C, Weinstein P, Bi P (2017) Environ Res 153:17–26

Xu J, Xu K, Li Z, Meng F, Tu T, Xu L, Liu Q (2020) Int J Environ Res Public Health 17

Yousef R, Khan S, Gupta G, Siddiqui T, Albahlal BM, Alajlan SA, Haq MA (2023) Diagnostics (Basel) 13

Zhao N, Charland K, Carabali M, Nsoesie EO, Maheu-Giroux M, Rees E, Yuan M, Garcia Balaguera C, Jaramillo Ramirez G, Zinszer K (2020) PLoS Negl Trop Dis 14:e0008056