

Prognosis of vector borne dengue disease outbreak in urban areas using multivariate analysis

Pratik S. Machchar¹, Purvi N. Ramanuj², Rajan Patel³, Jitendra Bhatia⁴, Kuntesh Jani⁵

¹Gujarat Technological University, Ahmedabad, India

²Vishwakarma Government Engineering College, Ahmedabad, India

³Kalol Institute of Technology and Research Center, India

⁴Nirma University, Ahmedabad, India

⁵L.D. College of Engineering, Ahmedabad, India

Article Info

Article history:

Received August 22, 2025

Revised March 14, 2026

Accepted May 21, 2026

Keywords:

Deep learning

Dengue fever

Machine learning

Time series forecasting

Vector-borne diseases

ABSTRACT

Vector borne disease like dengue continues to pose a significant climate-sensitive public health challenge in tropical regions such as Brazil, Peru, and India. This study examines the feasibility of predicting dengue outbreaks using weekly multivariate time-series data from San Juan (SJ), Puerto Rico and Iquitos (IQ), Peru. Dengue incidence was analyzed alongside meteorological, environmental, and vegetation-based variables to capture key climatic influences. Several machine learning and deep learning approaches were evaluated, including LightGBM. Model performance was assessed using root mean square error (RMSE) and mean absolute error (MAE). The results show that LightGBM achieved the lowest RMSE/MAE, indicating strong short-term predictive accuracy and excellent interpretability. Feature importance analysis and principal component analysis (PCA) identified precipitation, dew point temperature, and humidity as the most influential predictors of dengue incidence. The study demonstrates that advanced machine learning models can serve as reliable early warning systems for vector-borne diseases. While this research focuses on dengue, the methodology is adaptable to other vector-borne datasets and diseases, offering a flexible tool for public health authorities to predict and mitigate outbreaks in diverse urban contexts.

This is an open access article under the [CC BY-SA](#) license.



Corresponding Author:

Pratik S. Machchar

Gujarat Technological University

Ahmedabad, Gujarat, India

Email: pratikmachchar@gmail.com

1. INTRODUCTION

Dengue fever continues to pose a significant and growing public health threat in tropical and subtropical regions worldwide. Transmitted primarily by *Aedes aegypti* and *Aedes albopictus* mosquitoes, dengue now places nearly half of the global population at risk of infection. In 2024, the World Health Organization (WHO) reported over 14.4 million dengue cases and more than 11,200 deaths across all six WHO regions, marking one of the highest annual global burdens on record [1]. In India, national surveillance data indicate a substantial increase in dengue incidence, with over 233,000 reported cases and nearly 300 associated deaths in 2024, continuing a trend of high disease burden in recent years [1]. These figures reflect the expanding geographic spread and climate-sensitive dynamics of dengue transmission, underscoring its persistent public health impact.

The four dengue serotypes (DENV1-4) cause illness ranging from mild fever to life-threatening hemorrhagic complications [2]. Urbanization and climate change have expanded mosquito habitats. These vectors also transmit malaria, Zika, chikungunya, and yellow fever. Environmental factors like temperature, rainfall, and humidity affect mosquito breeding and virus replication [3]-[5]. Higher temperatures accelerate virus development, rainfall creates breeding sites [6], [7], and vegetation density provides ideal mosquito habitats [8].

Despite advances in surveillance, predicting dengue outbreaks remains difficult [9]-[11]. Existing methods struggle to capture interactions between climate, human behavior, and mosquito dynamics [12], limiting timely public health responses [13]-[15].

This study evaluates multivariate time series forecasting for dengue prediction using data from San Juan (SJ) and Iquitos (IQ). Most previous work focuses on univariate statistical/deep learning methods, or often with limited predictors [9], [16], [17]. In this study, six multivariate models —SARIMAX, LightGBM, Seq2Seq, RNN, CNN-BiLSTM, and Stacked LSTM— comparing statistical, machine learning, and deep learning approaches under identical conditions [11], [18]-[26].

For each model, principal component analysis (PCA) identifies key climate drivers of dengue incidence, addressing the interpretability gap in ensemble methods [27], [28]. Weekly forecasts align with operational surveillance needs, unlike annual or long-term studies [18], [29].

This work contributes:

- An integrated framework combining climate and epidemiological variables.
- Systematic cross-paradigm model evaluation comparing classical statistical methods (SARIMAX), machine learning approaches (LightGBM), and deep learning architectures (Seq2Seq, RNN, CNN-BiLSTM, Stacked LSTM) under identical experimental conditions to identify the most effective forecasting approach
- Identification of influential predictors/variables that contribute most to dengue incidence.
- Comparison of weekly forecast ability for practical disease surveillance.

The results show that LightGBM outperforms other approaches, with precipitation, dew point temperature, and humidity identified as the most influential predictors. While demonstrated using dengue, the framework is broadly applicable to other vector-borne diseases in urban settings.

2. MATERIALS AND METHODS

2.1. Dataset

The dataset used for this study focuses on vector-borne disease predictions, such as dengue, and includes detailed climate and environmental data for two cities: SJ and IQ [30] (Figure 1).

It spans several years on a weekly timescale, capturing the correlation between prevailing climate variables and dengue outbreaks. Key features include dengue case counts alongside various climate metrics such as temperature, precipitation, and humidity, all of which are known to influence the transmission of vector-borne diseases.

2.2. Data preprocessing

Effective data preprocessing is crucial for improving the accuracy of models predicting vector-borne disease outbreaks. In this study, multiple steps were taken to clean and prepare the dataset before applying forecasting models.

- Data formatting: The date column was converted into a standard DateTime format. Abnormal or erroneous dates (e.g., 1994-02-31) were corrected by replacing them with corresponding dates from the previous year to maintain temporal consistency.
- Handling missing and incorrect values: Incorrect or false values, including null values, were systematically identified and removed from the dataset to prevent distortion in model predictions.
- Outlier detection and replacement: Outliers were detected based on statistical thresholds and replaced with NaN values to prevent skewed data points from influencing the analysis.
- Interpolation missing values, including those introduced by outlier replacement, were filled using interpolation methods, ensuring a continuous dataset without introducing bias.
- Data scaling: To standardize the dataset and avoid any bias due to differing value ranges across features, data scaling was applied, ensuring uniformity across all variables for model training.

These preprocessing steps were essential for refining the dataset, enabling accurate and reliable dengue outbreak predictions using various forecasting models. The dataset incorporates weather station data from

NOAA, such as maximum, minimum, and average temperatures, total precipitation, and diurnal temperature range. Additionally, reanalysis data includes relative and specific humidity, air temperature, dew point temperature, and precipitation, providing a comprehensive understanding of climatic factors. Satellite-based precipitation (PERSIANN) and vegetation indices (NDVI) offer further insights into environmental conditions. These features enable a multivariate analysis of how factors like temperature, rainfall, and humidity impact dengue transmission dynamics in urban areas.

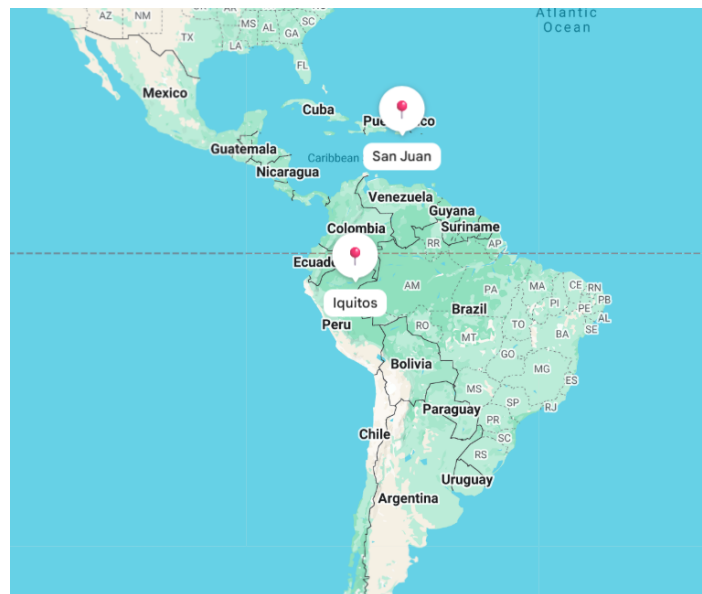


Figure 1. Study locations: San Juan (Puerto Rico) and Iquitos (Peru)

2.3. Available parameters for predicting dengue cases

The dataset comprises 24 environmental and climatic features alongside the target variable (total dengue cases), providing a comprehensive multivariate framework for outbreak prediction. Before analyzing the impact of individual climate parameters, it is important to understand how each variable contributes to dengue transmission. The following sections detail the main climate and environmental features considered in this study, starting with temperature, which plays a crucial role in mosquito development and virus replication.

2.3.1. Temperature

- station_max_temp_c, station_min_temp_c, station_avg_temp_c: Weather station data capturing daily temperature fluctuations, critical because *Aedes* mosquitoes thrive within certain temperature ranges. Higher temperatures speed up mosquito life cycles and virus replication.
- reanalysis_max_air_temp_k, reanalysis_min_air_temp_k, reanalysis_avg_temp_k: Reanalysis data provides broader scale temperature insights, important for capturing regional climate patterns that affect mosquito habitats.
- reanalysis_tdr_k, station_diur_temp_rng_c: Diurnal temperature range indicates the difference between day and night temperatures. This range affects mosquito activity, biting behavior, and survival, directly influencing dengue transmission rates.

2.3.2. Precipitation

- station_precip_mm, precipitation_amt_mm, reanalysis_precip_amt_kg_per_m2, reanalysis_sat_precip_amt_mm: Precipitation creates breeding grounds for *Aedes* mosquitoes, particularly in stagnant water. The amount of rainfall influences mosquito population density, while heavy rains may disrupt breeding cycles by washing away larvae. Monitoring precipitation is essential for understanding vector proliferation and disease risk.

2.3.3. Humidity

- `reanalysis_relative_humidity_percent`, `reanalysis_specific_humidity_g_per_kg`: Humidity plays a pivotal role in mosquito survival and longevity. Higher humidity increases the lifespan of mosquitoes, allowing them to bite multiple humans and thus spread the disease, while relative humidity measures the percentage of saturation.
- `reanalysis_dew_point_temp_k`: Dew point temperature reflects moisture levels in the atmosphere. Higher dew points create favorable conditions for mosquito survival and virus transmission.

2.3.4. Diurnal temperature range

- `station_diur_temp_rng_c`: A smaller diurnal temperature range indicates stable temperatures, which favor consistent mosquito activity. Wider ranges may disrupt mosquito activity but can also shorten the virus incubation period within the vector.

2.3.5. Vegetation index (NDVI)

- `ndvi_se`, `ndvi_sw`, `ndvi_ne`, `ndvi_nw`: NDVI measures vegetation density around the city. Dense vegetation provides ideal breeding and resting places for mosquitoes. Higher NDVI values generally correspond to higher mosquito populations, increasing the risk of vector-borne disease transmission in those areas. Monitoring vegetation helps identify hotspots for vector activity.

Each of these parameters offers unique insights into the environmental and climatic conditions conducive to dengue transmission. Their combined analysis enables more accurate predictions of vector-borne disease outbreaks.

2.4. Principal component analysis

Since the environmental, epidemiological, and meteorological features are highly correlated, PCA was conducted to reduce dimensionality, thereby improving model stability and prediction performance [8], [12], [21]. The dataset initially contained 24 climatic, environmental, and epidemiological features. PCA was conducted to reduce dimensionality and identify the most informative predictors. As shown in Figure 2, through multiple trials and experiments, four key features were finalized: precipitation amount (`precipitation_amt_mm`), dew point temperature (`reanalysis_dew_point_temp_k`), specific humidity (`reanalysis_specific_humidity_g_per_kg`) and historical total cases (`total_cases`). These features consistently demonstrated the strongest contribution to explaining variance in dengue incidence patterns. The use of these four features across different models is shown in Figure 2.

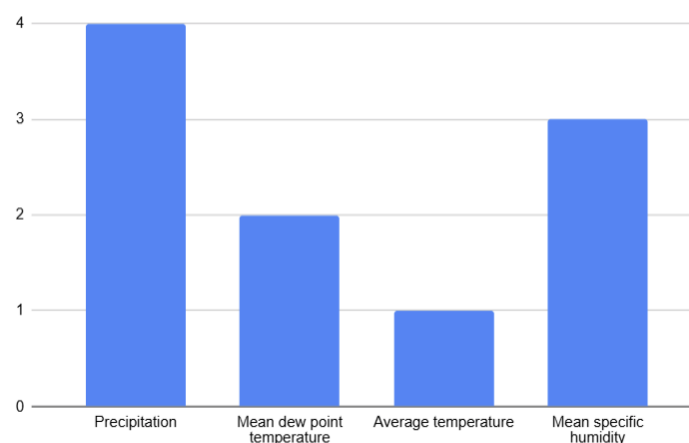


Figure 2. PCA analysis showing feature importance across all models

2.5. Hyperparameter tuning

Hyperparameter tuning methods like Bayesian [13] and grid search [29] are popular methods for hyperparameter optimization. All models were fine-tuned using Optuna [13]. For each model, a predefined search space was explored over multiple trials, where model performance was evaluated using validation loss. The

tuning process was carried out in a consistent manner across models to identify stable parameter settings. The best-performing hyperparameter configurations were then selected and used for the final training and evaluation of each model. Details of the optimized hyperparameters for each model are provided in Appendix A.

3. RESULTS AND DISCUSSION

In this study, multiple machine learning and deep learning models were evaluated to forecast dengue outbreaks using multivariate time series data from San Juan (sj) and Iquitos (iq). Six models were tested, including LightGBM, Stacked LSTM, CNN-BiLSTM, Seq2Seq, RNN, SARIMAX as shown in Table 1. The performance was assessed using root mean squared error (RMSE) and mean absolute error (MAE) metrics. RMSE was selected as the primary metric because it heavily penalizes large prediction errors, which is critical for dengue forecasting where missing outbreak spikes has serious public health consequences. MAE complements RMSE by providing the average typical error magnitude, and together these metrics enable assessment of both worst-case scenarios and consistent performance—essential for epidemiological forecasting where outbreak timing and magnitude are paramount. R^2 was excluded as it can be misleadingly high in time series with strong seasonal patterns and does not adequately penalize temporal lag or phase shifts in outbreak predictions. Based on comprehensive evaluation, LightGBM achieved the best performance with an RMSE of 16.94 and MAE of 11.42. This section presents an in-depth analysis of this model, while detailed results of all other models are provided in Appendix B for reference.

Table 1. Performance comparison of all models

Model	RMSE	MAE
LightGBM	16.94	11.42
Stacked LSTM	18.75	18.01
CNN-BiLSTM	22.13	19.38
Seq2Seq	22.23	13.44
RNN	25.27	13.98
SARIMAX	34.90	18.37

3.1. LightGBM-best model

The LightGBM model achieved the best performance with an RMSE of 16.94 and MAE of 11.42, demonstrating the effectiveness of gradient boosting for dengue forecasting. The model showed strong predictive capability and excellent interpretability (see Figure 3).

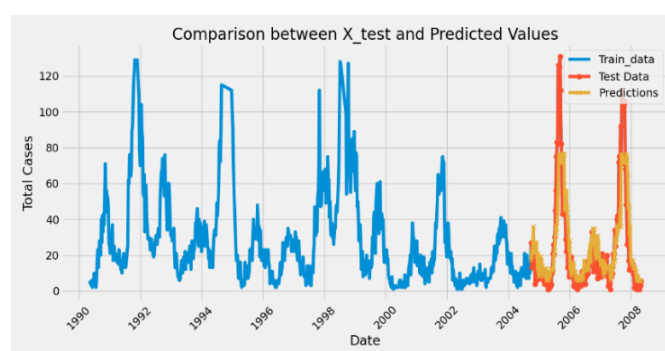


Figure 3. Actual vs predicted v for LightGBM model

3.1.1. Predictive performance analysis

The LightGBM model's predictions demonstrate strong alignment with actual outbreak patterns using a multivariate approach that integrates four key features: precipitation amount, dew point temperature, total cases, and specific humidity. The model excels at capturing overall outbreak trends with minimal noise in predictions and successfully predicts the magnitude of outbreak peaks, though occasionally underestimates

extreme events. The gradient boosting framework with 245 estimators effectively learns complex interactions between climate and epidemiological variables.

3.1.2. Feature importance analysis

The LightGBM model was trained using four key features selected through feature selection and PCA: precipitation amount (precipitation_amt_mm), specific humidity (reanalysis_specific_humidity_g_per_kg), dew point temperature (reanalysis_dew_point_temp_k), and historical total cases (total_cases). The model's feature importance analysis revealed that precipitation emerged as the dominant predictor among these variables, indicating that rainfall patterns create favorable breeding conditions for mosquito vectors. Dew point temperature and specific humidity, both moisture-related variables, significantly influenced predictions by affecting vector survival and reproduction rates. The tree-based structure of LightGBM provides excellent interpretability, allowing the model to capture non-linear interactions between these climate and epidemiological variables.

3.2. Comparative analysis

LightGBM, with an RMSE of 16.94 and MAE of 11.42, provides excellent interpretability and captures complex non-linear relationships effectively. LightGBM is suited for model interpretability and understanding of feature interactions, making it ideal for public health applications where transparency in decision-making is crucial.

3.3. Principal component analysis and feature selection across models

PCA was conducted to identify the most influential climate and environmental features for dengue prediction. The analysis revealed that three moisture-related variables precipitation (precipitation_amt_mm), dew point temperature (reanalysis_dew_point_temp_k) and specific humidity (reanalysis_specific_humidity_g_per_kg) collectively accounted for the majority of variance in dengue incidence, confirming their critical role in disease transmission dynamics.

Precipitation emerged as the most frequently selected feature, appearing in four of six models (LightGBM, RNN, Seq2Seq, and SARIMAX), confirming its fundamental role in dengue prediction. The best-performing models utilized comprehensive moisture-related variables-LightGBM (RMSE 16.94) incorporated precipitation, dew point temperature, and specific humidity alongside historical case data. While multivariate models generally outperformed univariate approaches, Stacked LSTM (RMSE 18.75) demonstrated competitive performance using only historical case patterns, indicating that temporal autocorrelation in disease incidence contains substantial predictive information. The PCA-guided feature selection strategy successfully identified influential predictors while reducing model complexity, providing strong evidence for climate-driven dengue transmission patterns.

3.4. Simplicity vs complexity trade-off

The deep learning models, including Seq2Seq, RNN, CNN-BiLSTM, and Stacked LSTM, provided moderate to good performance, with Stacked LSTM performing the best among them. However, the LightGBM model, which is relatively simpler in architecture, outperformed the deep learning models. This outcome challenges the assumption that deeper, more complex architectures necessarily yield better results in multivariate time series forecasting tasks. The superior performance of traditional machine learning approaches suggests that for the given dataset size and feature space, model complexity does not always translate to improved prediction accuracy.

3.5. LightGBM's superior performance

interpretability through its tree-based structure and gradient boosting framework. Its ability to provide clear feature importance rankings and capture non-linear interactions between climate variables makes it valuable for understanding the drivers of dengue outbreaks. Its transparent decision-making process is crucial for public health applications where understanding model predictions is as important as accuracy. The model's computational efficiency and strong predictive capability make it ideal for operational forecasting and strategic planning.

3.6. Deep learning limitations

The moderate performance of deep learning models (Stacked LSTM, CNN-BiLSTM, Seq2Seq, and RNN) compared to LightGBM highlights potential limitations when applied to this type of dengue dataset.

While these architectures excel at capturing long-term dependencies and complex temporal patterns, the characteristics of dengue outbreak data—including irregular periodicity, sparse extreme events, and climate-disease interaction complexities—may pose challenges that traditional machine learning approaches handle more efficiently. The computational cost and relatively higher RMSE suggest that deep learning models need further refinement or additional training data to compete with traditional approaches in this forecasting context.

3.7. Classical model constraints

The SARIMAX model's poor performance underscores the limitations of classical time series models when applied to multivariate forecasting tasks. SARIMAX's statistical framework, while effective for univariate analysis, struggles to adequately model the complex interactions between multiple climate variables that influence dengue outbreaks. This limitation emphasizes the need for machine learning approaches that can capture non-linear relationships and feature interactions in epidemiological forecasting.

The findings have practical implications for dengue surveillance systems. The models provide a critical timeframe for implementing preventive interventions and resource allocation. Feature importance analysis consistently identified precipitation, dew point temperature, and humidity as key predictors, which can guide targeted mosquito control efforts during high-risk climatic conditions. LightGBM is ideal for both real-time forecasting and strategic planning decisions due to its balance of accuracy and interpretability.

Current limitations include the dataset being restricted to San Juan (SJ) and Iquitos (Iq), focus on climate variables without socioeconomic factors, and weekly resolution that may miss finer temporal dynamics. Future research should explore ensemble approaches combining model strengths, integration of additional data sources such as satellite imagery and mobility data, development of city-specific models, and investigation of climate change scenarios for long-term outbreak predictions.

Note: Detailed results for all other evaluated models (Seq2Seq, RNN, CNN-BiLSTM, Stacked LSTM, and SARIMAX) are provided in Appendix B, including performance metrics, visualizations, and model-specific analyses.

4. CONCLUSION

This research demonstrates the effectiveness of using multivariate analysis and time series forecasting techniques for predicting outbreaks of vector-borne diseases, with a specific focus on dengue as a demonstration case. The models explored in this study, including Seq2Seq, RNN, CNN-BiLSTM, Stacked LSTM, and LightGBM. Provides a diverse set of approaches for capturing the complex interactions between environmental factors and disease transmission. The results highlight the potential of advanced deep learning models, such as Stacked LSTM and CNN-BiLSTM, and machine learning models like LightGBM SARIMAX, which struggled to handle the dataset's complexity. These findings align with recent comparative studies demonstrating the superiority of machine learning approaches for dengue forecasting. These deep learning models were able to model the intricate temporal and seasonal patterns that are critical in disease forecasting, thus offering more accurate predictions. The study reinforces that using a range of models—spanning from deep learning to statistical approaches—offers a comprehensive method for forecasting vector-borne disease outbreaks, enabling public health authorities to deploy proactive measures to control the spread of disease in urban areas.

5. FUTURE WORK

The future work of this research aims to enhance the model's capabilities by exploring more advanced models and comparing their performance in predicting vector-borne diseases. Additionally, expanding the dataset to include diverse geographical regions and other diseases like malaria, chikungunya, and Zika will allow the model to generalize better. Future research may also suggest new, improved models while incorporating additional variables, such as socio-economic and environmental factors. Furthermore, integrating real-time data streams could improve prediction accuracy and assist in the development of early warning systems for outbreak mitigation. Large language models may also be explored for enhanced forecasting capabilities.

FUNDING INFORMATION

Authors state that no funding was involved.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Pratik S. Machchar	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓	
Purvi N. Ramanuj		✓										✓		
Rajan Patel										✓				
Jitendra Bhatia										✓				
Kuntesh Jani										✓				

C : Conceptualization

M : Methodology

So : Software

Va : Validation

Fo : Formal Analysis

I : Investigation

R : Resources

D : Data Curation

O : Writing - Original Draft

E : Writing - Review & Editing

Vi : Visualization

Su : Supervision

P : Project Administration

Fu : Funding Acquisition

CONFLICT OF INTEREST STATEMENT

Authors declare no conflict of interest.

DATA AVAILABILITY

Derived data supporting the findings of this study, including dengue case records and climate variables, and the code used for model development is available from the author [P.S.M.] upon request. Some data may be subject to institutional or privacy restrictions.

REFERENCES

- [1] World Health Organization (WHO), "Dengue - Global situation," *Weekly Epidemiological Record*, vol. 100, no. 52, pp. 665–678, 2024, [Online]. Available: <https://www.who.int/publications/i/item/who-wer10052-665-678>.
- [2] P. Silitonga, B. Dewi, A. Bustamam, and H. Al-Ash, "Evaluation of dengue model performances developed using artificial neural network and random forest classifiers," *Procedia Computer Science*, vol. 179, pp. 135–143, 2021, doi: 10.1016/j.procs.2020.12.018.
- [3] C. Caminade *et al.*, "Impact of recent and future climate change on vector-borne diseases," *Annals of the New York Academy of Sciences*, vol. 1436, no. 1, pp. 157–173, 2019, doi: 10.1111/nyas.13950.
- [4] P. Arcari, N. Tapper, and S. Pfueller, "Regional variability in relationships between climate and dengue/DHF in Indonesia," *Singapore Journal of Tropical Geography*, vol. 28, pp. 251–272, 2007, doi: 10.1111/j.1467-9493.2007.00300.x.
- [5] Y. L. Hii, H. Zhu, N. Ng, L. C. Ng, and J. Rocklöv, "Forecast of dengue incidence using temperature and rainfall," *PLOS Neglected Tropical Diseases*, vol. 6, no. 11, p. e1908, 2012, doi: 10.1371/journal.pntd.0001908.
- [6] S. Bal and S. Sodoudi, "Modeling and prediction of dengue occurrences in Kolkata, India, based on climate factors," *International Journal of Biometeorology*, vol. 64, no. 8, pp. 1379–1391, 2020, doi: 10.1007/s00484-020-01918-9.
- [7] M. Gharbi *et al.*, "Time series analysis of dengue incidence in Guadeloupe, French West Indies: forecasting models using climate variables as predictors," *BMC Infectious Diseases*, vol. 11, p. 166, 2011, doi: 10.1186/1471-2334-11-166.
- [8] Á. G. Muñoz *et al.*, "AeDES: a next-generation monitoring and forecasting system for environmental suitability of Aedes-borne disease transmission," *Scientific Reports*, vol. 10, no. 1, p. 12640, 2020, doi: 10.1038/s41598-020-69625-4.
- [9] S. S. Deshmukh and S. Deb, "A survey of statistical and machine learning methods of quantile regression in time series and their suitability in predicting dengue outbreaks," *Japanese Journal of Statistics and Data Science*, pp. 1–49, 2025, doi: 10.1007/s42081-025-00297-y.
- [10] S. Gambhir, S. K. Malik, and Y. Kumar, "The diagnosis of dengue disease: an evaluation of three machine learning approaches," *International Journal of Healthcare Information Systems and Informatics*, vol. 13, no. 3, pp. 1–19, 2018.
- [11] P. Siritasatien, S. Chadsuthi, K. Jampachaisri, and K. Kesorn, "Dengue epidemics prediction: A survey of the state-of-the-art based on data science processes," *IEEE Access*, vol. 6, pp. 53757–53795, 2018, doi: 10.1109/ACCESS.2018.2871241.
- [12] R. Jain, S. Sontisirikit, S. Iamsirithaworn, and H. Prendering, "Prediction of dengue outbreaks based on disease surveillance, meteorological and socio-economic data," *BMC Infectious Diseases*, vol. 19, no. 1, p. 272, 2019, doi: 10.1186/s12879-019-3874-x.
- [13] S. G. Kakarla *et al.*, "Weather integrated multiple machine learning models for prediction of dengue prevalence in India," *International Journal of Biometeorology*, vol. 67, no. 2, pp. 285–297, 2023, doi: 10.1007/s00484-022-02405-z.
- [14] W. Anggraeni *et al.*, "Prediction of dengue fever outbreak based on climate factors using fuzzy-logistic regression," in *Proc. 2020 International Seminar on Intelligent Technology and Its Applications (ISITIA)*, 2020, pp. 199–204, doi: 10.1109/ISITIA49792.2020.9163708.
- [15] O. S. Baquero, L. M. R. Santana, and F. Chiaravalloti-Neto, "Dengue forecasting in São Paulo city with generalized additive models, artificial neural networks and seasonal autoregressive integrated moving average models," *PLoS One*, vol. 13, no. 4, p. e0195065, 2018, doi: 10.1371/journal.pone.0195065.

- [16] P. Guo *et al.*, “Developing a dengue forecast model using machine learning: a case study in China,” *PLOS Neglected Tropical Diseases*, vol. 11, no. 10, p. e0005973, 2017, doi: 10.1371/journal.pntd.0005973.
- [17] J. Xu *et al.*, “Forecast of dengue cases in 20 Chinese cities based on the deep learning method,” *International Journal of Environmental Research and Public Health*, vol. 17, no. 2, p. 453, 2020, doi: 10.3390/ijerph17020453.
- [18] E. Mussumeci and F. C. Coelho, “Large-scale multivariate forecasting models for Dengue - LSTM versus random forest regression,” *Spatial and Spatio-temporal Epidemiology*, vol. 35, p. 100372, 2019, doi: 10.1016/j.sste.2020.100372.
- [19] A. L. Ramadona, L. Lazuardi, Y. L. Hii, A. Holmner, H. Kusnanto, and J. Rocklöv, “Prediction of dengue outbreaks based on disease surveillance and meteorological data,” *PLoS One*, vol. 11, no. 3, p. e0152688, 2016, doi: 10.1371/journal.pone.0152688.
- [20] T. K. Yamana, S. Kandula, and J. Shaman, “Superensemble forecasts of dengue outbreaks,” *Journal of the Royal Society Interface*, vol. 13, no. 123, p. 20160410, 2016, doi: 10.1098/rsif.2016.0410.
- [21] M. A. Mobin, “Multivariate forecasting of dengue infection in Bangladesh: evaluating the influence of data downscaling on machine learning predictive accuracy,” *BMC Infectious Diseases*, vol. 25, no. 1, p. 761, 2025, doi: 10.1186/s12879-025-11159-z.
- [22] X. Chen and P. Moraga, “Assessing dengue forecasting methods: a comparative study of statistical models and machine learning techniques in Rio de Janeiro, Brazil,” *Tropical Medicine and Health*, vol. 53, no. 1, p. 52, 2025, doi: 10.1186/s41182-025-00723-7.
- [23] A. E. Laureano-Rosario *et al.*, “Application of artificial neural networks for dengue fever outbreak predictions in the northwest coast of Yucatan, Mexico and San Juan, Puerto Rico,” *Tropical Medicine and Infectious Disease*, vol. 3, no. 1, p. 5, 2018, doi: 10.3390/tropicalmed3010005.
- [24] I. Sutskever, O. Vinyals, and Q. V. Le, “Sequence to sequence learning with neural networks,” *arXiv*, 2014, [Online]. Available: <https://arxiv.org/abs/1409.3215>.
- [25] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural Computation*, vol. 9, pp. 1735–1780, 1997, doi: 10.1162/neco.1997.9.8.1735.
- [26] G. Ke *et al.*, “LightGBM: a highly efficient gradient boosting decision tree,” in *Proc. 31st International Conference on Neural Information Processing Systems (NIPS'17)*, 2017, pp. 3149–3157.
- [27] M. Cabrera *et al.*, “Dengue prediction in Latin America Using machine learning and the one health perspective: A literature review,” *Tropical Medicine and Infectious Disease*, vol. 7, p. 322, 2022, doi: 10.3390/tropicalmed7100322.
- [28] H. Barlevi, Z. E. R. Cordero, F. J. Colon-Gonzalez, and R. Lowe, “A reproducible ensemble machine learning approach to forecast dengue outbreaks,” *Scientific Reports*, vol. 14, no. 1, p. 3807, 2024, doi: 10.1038/s41598-024-52796-9.
- [29] Z. I. Lai, W. K. Fung, and E. Chew, “Machine learning models for dengue forecasting in Singapore,” *arXiv preprint arXiv:2407.00332*, 2024, doi: 10.48550/arXiv.2407.00332.
- [30] Y. Qian, “DengAI Dataset,” *Kaggle*, [Online]. Available: <https://www.kaggle.com/datasets/qianyigang129/dengai-dataset>.

APPENDICES

APPENDIX A. Model hyperparameters

This subsection summarizes the main hyperparameters and settings for each model as used in this study.

A.1. LightGBM

Input features: [precipitation_amt_mm, reanalysis_dew_point_temp_k, total_cases, reanalysis_specific_humidity_g_per_kg] (Multivariate). Learning rate: 0.00475, L1 reg: 0.0981, L2 reg: 0.0141, Num leaves: 245, Num estimators: 245, Feature fraction: 0.8150, Max depth: 8, Min data in leaf: 74, RMSE: 16.94.

A.2. Stacked LSTM

Input features: [total_cases] (Univariate), Kernel initializer: He normal, Number of layers: 1, Neurons: 69, Dropout rate: 0.105, Epochs: 97, Batch size: 16, Optimizer: Adam, Activation function: ReLU, Early stopping patience: 16, RMSE: 18.75.

A.3. CNN-BiLSTM

Input features: [total_cases] (Univariate), Number of convolutional filters: 39, Kernel size: 3, BiLSTM neurons: 127, Epochs: 19, Batch : 16, Optimizer: SGD, Activation: ELU, Early stopping patience: 11, RMSE: 22.13.

A.4. Seq2Seq

Input features: [precipitation_amt_mm, reanalysis_specific_humidity_g_per_kg, total_cases] (Multivariate), Window length: 8, Neurons: 47, Dropout rate: 0.174, Epochs: 61, Batch size: 5, Optimizer: Nadam, Early stopping patience: 17, RMSE: 22.23.

A.5. RNN

Input features: [precipitation_amt_mm, reanalysis_dew_point_temp_k, total_cases, reanalysis_specific_humidity_g_per_kg] (Multivariate). Kernel init: Glorot uniform, Window length: 5, Num layers: 1, Neurons: 48, Dropout: 0.111, Epochs: 47, Batch size: 10, Optimizer: RMSprop, Activation: ReLU, Early stopping patience: 14, RMSE: 25.27.

A.6. SARIMAX

Input (exogenous) features: [precipitation_amt_mm, reanalysis_avg_temp_k], Target variable: total_cases, Type: Multivariate, Forecasting horizon: 4 steps ahead, RMSE: 34.90.

APPENDIX B. Detailed results of other models

This subsection provides comprehensive results for the remaining five models evaluated in the study: Seq2Seq, RNN, CNN-BiLSTM, Stacked LSTM, and SARIMAX.

B.1. Seq2Seq model

The Seq2Seq multi output model achieved an RMSE of 22.23. This model's moderate performance indicates its capability to capture sequential dependencies, though it did not perform optimally compared to the best models. See Figure 4 the Seq2Seq model shows general alignment with observed outbreaks but exhibits significant fluctuations during sharp changes.

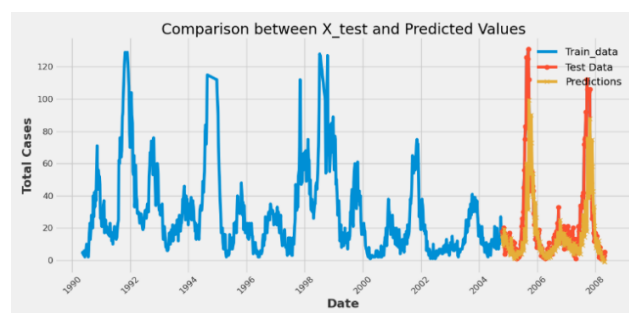


Figure 4. Actual vs predicted graph of Seq2Seq model

B.2. RNN model

The RNN multioutput model produced an RMSE of 25.27. The shallow structure with 1 layer and 48 neurons limited its ability to fully capture the nuances of the multivariate data. See Figure 5 the RNN model shows significant deviation from actual data during periods of rapid change, failing to predict higher peaks accurately.

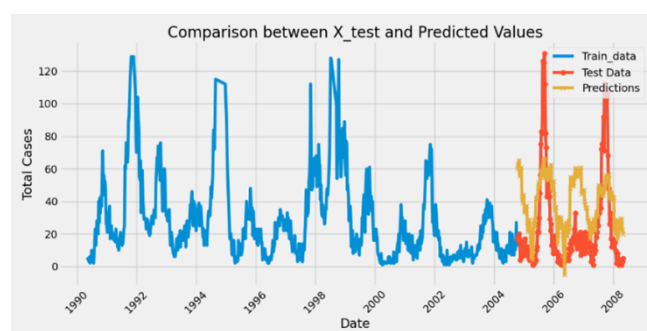


Figure 5. Actual vs predicted graph of RNN model

B.3. CNN-BiLSTM model

The CNN-BiLSTM model achieved an RMSE of 22.13, combining convolutional layers for spatial feature extraction with bidirectional LSTMs for temporal dependencies. See Figure 6 the CNN-BiLSTM shows reasonable alignment with actual data though struggles with certain peaks.

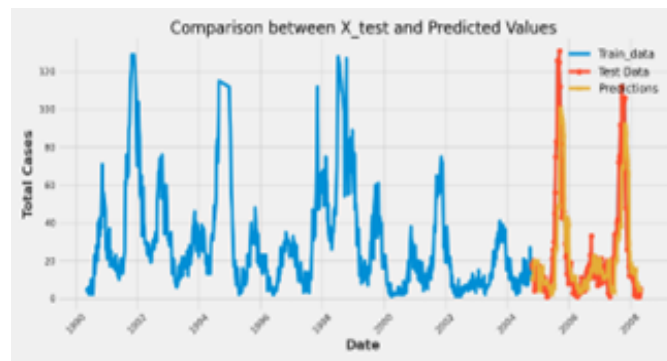


Figure 6. Actual vs predicted graph of CNN-BiLSTM model

B.4. Stacked LSTM model

The Stacked LSTM model achieved an RMSE of 18.75, ranking as the best among deep learning models. Its multiple LSTM layers enable learning of both short and long-term temporal dependencies. See Figure 7 the Stacked LSTM demonstrates good ability to model overall trends with close alignment for most periods.

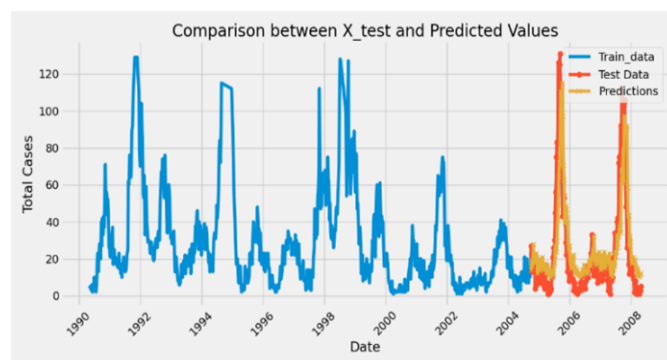


Figure 7. Actual vs predicted graph of stacked LSTM model

B.5. SARIMAX model

The SARIMAX model produced the highest RMSE of 34.90, highlighting limitations of classical time series models for multivariate forecasting with complex temporal and spatial interactions. See Figure 8 the SARIMAX model shows the largest discrepancy between predicted and actual cases, struggling to track volatility in the dataset.

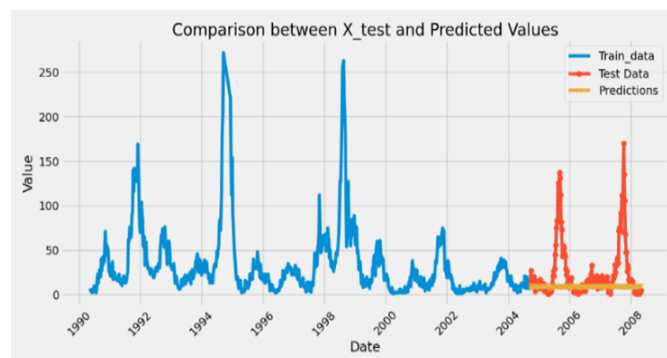


Figure 8. Actual vs predicted graph of SARIMAX model

BIOGRAPHIES OF AUTHORS

Pratik S. Machchar    is a research scholar at Gujarat Technological University, Ahmedabad, India. His research interests include practical implementation of time series forecasting, and epidemiological modeling. His current work focuses on developing predictive models for vector-borne disease outbreaks using advanced computational techniques. He has contributed to this study through model development, data analysis, and manuscript preparation. He can be contacted at email: pratikmachchar@gmail.com.



Dr. Purvi N. Ramanuj    is affiliated with Vishwakarma Government Engineering College, Ahmedabad, India. She specializes in computer science and engineering with research interests in artificial intelligence, data science, and computational modeling. Her expertise has contributed significantly to the methodological framework and validation processes in this study. She can be contacted at email: purviramanuj@yahoo.com.



Dr. Rajan Patel    is associated with Kalol Institute of Technology and Research Center, India. His research areas include machine learning applications in healthcare, predictive modeling, and computational intelligence. He has contributed to this research through technical guidance and conceptual development of the forecasting framework. He can be contacted at email: rg-pce21@gmail.com.



Dr. Jitendra Bhatia    is affiliated with Nirma University, Ahmedabad, India. His expertise encompasses artificial intelligence, deep learning architectures, and their applications in healthcare informatics. His contributions to this study include critical review of model architectures and performance evaluation strategies. He can be contacted at email: jitendrabbhatia@gmail.com.



Dr. Kuntesh Jani    is associated with L.D. College of Engineering, Ahmedabad, India. His research interests focus on data analytics, computational modeling, and application of machine learning techniques in epidemiological studies. He has provided valuable insights into the statistical analysis and interpretation of results in this research. He can be contacted at email: kuntesh.jani@ldce.ac.in.