Journal of Information Systems Engineering and Management
2025, 10(34s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Evaluating Feature Selection Techniques for Dengue Prediction with LSTM Model in Gujarat, India

Amiben Maheshbhai Mehta[1], Kajal Patel[2]
[1]*Research Scholar, Department of Computer Engineering, Gujarat Technological University, Ahmedabad, India.*
E-mail: mehtaamim@gmail.com
[2]*Associate Professor, Department of Computer Engineering, VGEC, Chandkheda, Gujarat Technological University, India*
E-mail: kajalpatel@vgecg.ac.in

ARTICLE INFO

Received: 31 Dec 2024

Revised: 20 Feb 2025

Accepted: 28 Feb 2025

ABSTRACT

Introduction: In some specific regions of Gujarat, India, dengue fever is a remarkable public health issue. Correct forecasting of dengue is required for its control and prevention. The proposed LSTM model, this research work investigates how feature selection methods—Correlation Coefficient, Recursive Feature Elimination (RFE), and Lasso Regression—affects dengue case forecasting improvement.

Objectives: The aim of this research study is to improve the predictive performance of the LSTM model by choosing the most relevant features, reducing computational complexity, and improving accuracy.

Methods: In proposed research study three feature selection methods are used to examine key characteristics including population density, climatic factors, and historical dengue cases. LSTM model is trained on chosen features, and its performance was assessed using RMSE and R^2 scores.

Results: It is observed that the best performance outcome from RFE-selected features, which had an RMSE of 0.05 and a R^2 score of 0.85. All features have the same RMSE but a lower R^2 score of 0.80, it suggests the power of feature selection.

Conclusions: dengue case prediction can be improved using LSTM feature selection. By increasing model interpretability and accuracy, RFE beat other techniques and underlined the requirement for optimal input variables in time-series forecasting for public health uses.

Keywords: Dengue Predictions, Feature Selection Techniques Public Health Management, LSTM Model, Gujarat

INTRODUCTION

The Thirteenth General Programme of Work 2019–2023 is a new five-year strategy plan that the World Health Organization (WHO) unveiled in January 2019 with the goal of enhancing the health and well-being of an extra billion people globally (World Health Organization, 2019a). Among the top 10 priority health challenges, dengue was identified as one of the four major infectious illnesses that pose a threat to world health (World Health Organization, 2019b). According to the WHO Global Strategy, diagnosis and case management, integrated surveillance and outbreak preparedness, sustainable vector control, future vaccine implementation, and foundational operational and implementation research are the five essential elements required to meet dengue public health targets. Climate variables including temperature, precipitation, and humidity are closely related to the spread of the Dengue virus (DENV) and have a big impact on how the disease outbreaks play out. Therefore, meteorological data analysis is essential for early epidemic detection and for comprehending environmental elements that either increase or reduce the development of disease. With 10,983 cases in 2021 and 18,219 cases in 2019—the greatest number between 2010 and 2022—the prevalence of dengue has been alarmingly on the rise in Gujarat, India. The highlight numbers indicate how urgently predictive solutions are needed to enable instant obtrusion and efficient dengue management.

Journal of Information Systems Engineering and Management
2025, 10(34s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Using a variety of datasets to produce useful insights the machine learning (ML) algorithms have represented their own values in disease forecasting or prediction across multiple locations. For example, a study carried out in Brazil which related with comparing different approaches for predicting dengue cases, including random forest regression, LASSO, and long short-term memory (LSTM) networks. The results indicated that LSTM performed better than the other methods in terms of prediction or forecasting the accuracy [1]. In Telangana, India, LSTM was used to estimate malaria incidence; in Ventakapuram the remarkable forecast accuracy of 96.11% was observed [2]. In addition to predicting diseases, LSTM models are proven effective and accurate in predicting weather, natural disasters, and air quality [3–4]. However, It is frequently necessary to choose pertinent input features for ML models to function well, and its impact on computational efficiency and prediction accuracy. By removing unwanted data and selecting very important variables the feature selection is incumbent for improving the model performance. Real-time applications are made possible by this procedure, which decreases computing complexity and overtraining hazards, while improving model interpretability [5]. In proposed research work we investigated how feature selection methods affects an LSTM-based model's ability to forecast dengue outbreaks in Gujarat, India. In order to pinpoint the most important environmental and epidemiological elements affecting dengue dynamics, here the researcher has specifically investigated three popular feature selection techniques like correlation coefficient analysis, recursive feature elimination (RFE), and LASSO regression. The motive behind the research study is to improve dengue management practices in the area and promote focused collective efforts, and increase prediction accuracy by implementing these tools.

DATA DESCRIPTION AND PRE-PROCESSING

The National Vector Borne Disease Control Programme (NVBDCP) for Gujarat, which includes the following six zones Ahmedabad, Gandhinagar, Vadodara, Surat, Rajkot, and Bhavnagar, provided the dengue case statistics. The weather and meteorological data such as monthly rainfall, humidity, and maximum and lowest temperatures are collected from India Meteorological Department (IMD).The IMD has provided center -specific meteorological data since district-specific data is not available. Ten districts of Gujarat states Ahmedabad, Amreli, Bhavnagar, Bhuj, Dwarka, Porbandar, Rajkot, Surat, Vadodara, and Valsad are selected for analysis in order to reconcile these statistics. Furthermore, population numbers and population density with effective from 2011 Indian Census are included. The combined data set includes 1200 rows and ten years of months monthly data from these areas which provides a strong foundation for prediction of dengue cases. To enable precise and accurate analysis, the pre-processing stage entailed standardizing the raw data and aligning the datasets geographically and chronologically. In order to assure the consistency across many data sources, this procedure is crucial. The creation of lag variables which represents the past values of particular variables at earlier time intervals is a key component of this phase. Lag variables have been developed for this investigation using one-month and twelve-month time lags for dengue and rainfall cases. With enabling the temporal examination of inter-variable connections, this approach is very helpful to identify patterns and trends that might not be immediately apparent. The dataset has normalized using the min-max scaling technique which converts the data into a range of 0 to 1, in order to get it ready for machine learning models. By assuring that the each feature contributes proportionately to the analysis this normalization step which improves the performance of predictive models.

FEATURE SELECTION TECHNIQUES

Effective feature selection strategies can be investigated by conducting a deep analysis of pertinent research studies in order to improve the accuracy of dengue case forecasts. The aim is to improve model performance without arbitrary feature expansion in order to guarantee significant and comprehensible outcomes. Many research studies demonstrate how well Random Forest and Support Vector Machine (SVM) works for electing features for regression analysis. The ability of Random Forest to handle the datasets with lots of variables is remarkable, especially when combined with SVM. It is observed and noted after performing experiments and data sets as per that the correlation coefficient (r-value) and the Root Mean Squared Error (RMSE) decreased consistently as per empirical results [6].By using feature selection methods such recursive feature elimination (RFE), Pearson's correlation coefficient and KBest analysis another study concentrated on differentiating dengue from other illnesses. With using this technique, the Critical features for a precise diagnosis of dengue are extracted from patient historical past data [7]. It has been noted that using ranking algorithms, Random Forest is found to be the

Journal of Information Systems Engineering and Management
2025, 10(34s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

best approach for rank based feature selection in medical data classification in a parallel study on disease prediction. This method proved very successful to forecast the illness like COVID-19 [8]. The research conducted using RFE and LASSO regularization feature selection techniques to forecast cervical cancer. The result shows that when these methods have combines with decision tree algorithms, the model performed exceptionally well [9]. A hybrid technique that used LASSO regression, SVM, and RFE have also successful to identify six distinctive genes as possible biomarkers for early diagnosis of pancreatic cancer research [10]. The Fast Correlation-Based Filter (FCBF) is applied in the context of unbalanced datasets for dengue infection cases in particular and a two-layer ensemble approach. This method comprises important pre-processing processes like class balance, noise reduction, outlier removal and the selection of important features, it is observed that all of these improvements increased the accuracy of predictions [11]. Finally, a logistic regression, RFE, and Fisher Score-based integrative feature selection technique called FRL is created for the purpose of identifying genetic biomarkers. Gene weights and dimensionality were first founded by Fisher Score and the ideal selection of features was then further optimized by RFE and logistic regression. The accuracy of this approach is higher than that of other feature selection methods [12]. Recursive Feature Elimination (RFE), Lasso Regression, and Pearson Correlation Coefficient Analysis are the feature selection methods used in this study after taking inspiration from prior studies. The accuracy and performance of the model has greatly improved by applying these techniques to the dataset gathered for dengue case prediction.

METHODOLOGY

In the research study titled "LSTM-based Forecasting of Dengue Cases in Gujarat: A Machine Learning Approach [13]," the LSTM model is implemented on the Gujarat dataset under three different conditions, as well as without employing any feature selection techniques. This approach utilises a fixed number of epochs as the stopping criterion. However, the current study throws lights on an improved methodology by incorporating feature selection techniques and an early stopping criterion. As depicted in Fig. 1, the data underwent collection and preprocessing steps. Following this, three feature selection techniques—recursive feature elimination (RFE), LASSO, and correlation coefficient analysis—were applied. The processed data was subsequently divided into training, validation, and testing sets, and the LSTM model was implemented for each feature selection technique. The training process for the model was governed by the early stopping criterion, where training ceased once the validation loss ceased to decrease, with a patience parameter set to 15 epochs. For each set of selected features derived from the various feature selection techniques, the model was retrained. Unlike the approach in [14], which relied on a fixed epoch count, early stopping effectively mitigates overfitting by terminating training when validation performance begins to degrade. This technique continuously monitors the mean squared error (MSE) loss on the validation set during training. If the validation loss stops decreasing and instead starts increasing over a specified number of epochs (patience), the training process is halted to avoid overfitting. It is important to note that the model obtained at the end of training may not always correspond to the best-performing model on the validation dataset. To address this, the model with the minimum validation loss during training is saved. This saved model, representing the optimal state, is subsequently tested on a separate test dataset, and its performance is evaluated using the root mean square error (RMSE) and R^2 score metrics.

Drawing from the literature survey conducted, this research implements three feature selection techniques: 1. Recursive Feature Elimination (RFE) 2. Lasso Regression 3. Pearson Correlation Coefficient Analysis

These techniques were applied to the following features: Average Temperature, Maximum Temperature, Minimum Temperature, Month of Year, Dengue cases with a one-month lag, Dengue cases with a lag of twelve months., Average Humidity, Total Rainfall with a Lag of One Month, Total Monthly Rainfall, Population Density, Population

The results obtained after applying these feature selection techniques are summarized in Table 1. The model development process employed a data partition ratio of 70% for training, 15% for validation, and 15% for testing. The Adam optimizer was utilized with a learning rate of 1e-3, a maximum of 150 epochs, a patience parameter set to 15, and a batch size of 4.

Journal of Information Systems Engineering and Management
2025, 10(34s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

PERFORMANCE METRICS

To test the efficacy of forecasting models the two widely recognized measures have employed which are root mean square error (RMSE) and the coefficient of determination (R^2). RMSE represents the accuracy of the model by means of a numerical measure of the prediction error measuring the discrepancy between observed and predicted values. In contrast the R^2 measures the proportion of variation in the dependent variable which is foreseeable from the independent variables and assessing the goodness of fit of the model. The evaluation method includes running the built model on an independent test dataset. This dataset has been used to forecast the dengue case count. The RMSE and R^2 values are used to evaluate the prediction performance of each created model later. These measurement readings show a full picture of the model's ability to generalize and its overall dependability in projecting dengue cases.

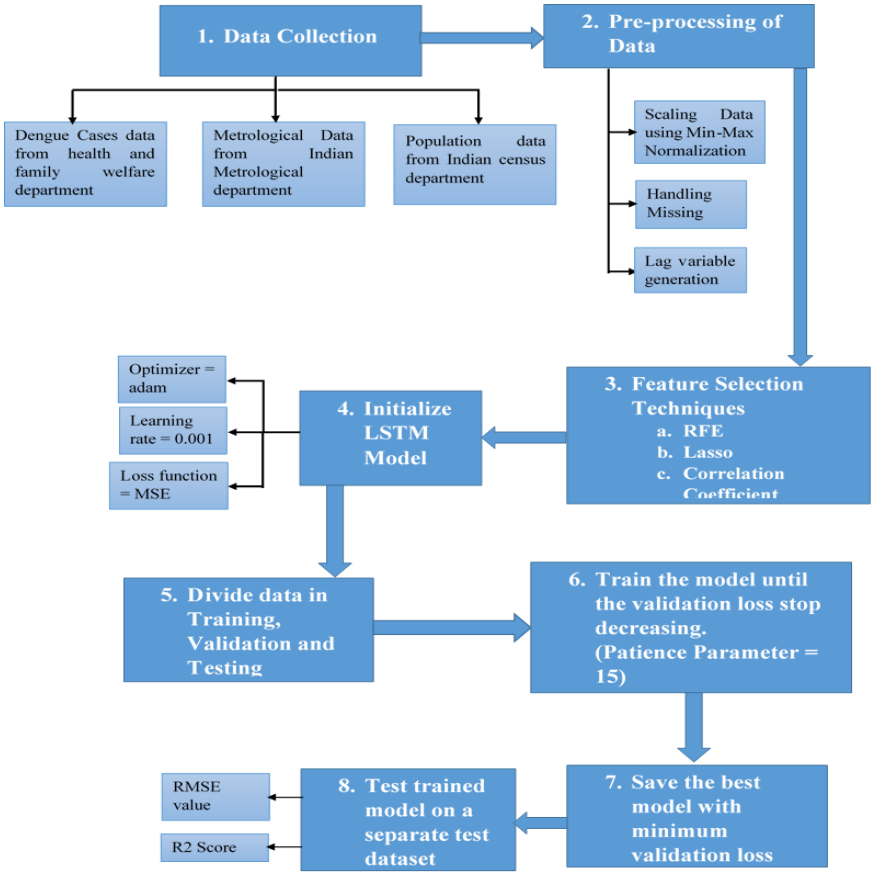

Figure 1. Proposed Approach for Dengue Case Predictions

Table 1. Result of Feature Selection Techniques

Feature Selection Technique	Features Extraction Result
RFE (Top 9)	1. Month 2. Maximum Temperature 3. Minimum Temperature 4. Average Temperature 5. Total Monthly Rainfall 6. Total Rainfall with lag of one month 7. Average Humidity 8. Dengue Case with lag of twelve month 9. Dengue Case with lag of one month
RFE(Top 8)	1. Maximum Temperature 2. Minimum Temperature 3. Average Temperature 4. Total Monthly Rainfall 5. Total Rainfall with lag of one month 6. Average Humidity 7. Dengue Case with lag of twelve month 8. Dengue Case with lag of one month
RFE(Top 7)	1. Maximum Temperature 2. Minimum Temperature 3. Average Temperature

Journal of Information Systems Engineering and Management
2025, 10(34s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

	4. Total Monthly Rainfall 5. Total Rainfall with lag of one month 6. Dengue Case with lag of twelve month 7. Dengue Case with lag of one month
Lasso	Descending Order: 1. Average Temperature 2. Maximum Temperature 3. Minimum Temperature 4. Month of Year 5. Dengue Case with lag of one month 6. Dengue Case with lag of twelve month 7. Average Humidity 8. Total Rainfall with lag of one month 9. Total Monthly Rainfall 10. Population Density 11. Population
Correlation Coefficient	Top features: 1. Dengue Case with lag of one month 2. Dengue Case with lag of twelve month 3. Population 4. Month of Year 5. Population Density 6. Total Rainfall with lag of one month 7. Average Humidity 8. Minimum Temperature 9. Total Monthly Rainfall 10. Average Temperature 11. Maximum Temperature

RESULT ANALYSIS

Table 2 summarizes the study of the selected feature set, highlighting the halting epoch at which the validation loss reaches its smallest value. It also indicates the Root Mean Squared Error (RMSE) values for training, testing, and validation datasets, along with the accompanying R^2 scores. The analysis shows that the Recursive Feature Elimination (RFE) technique achieves optimal performance when applied to the top eight features such as Maximum Temperature, Minimum Temperature, Average Temperature, Total Monthly Rainfall, Total Rainfall with a Lag of One Month, Average Humidity, Dengue Cases with a Lag of Twelve Months, and Dengue Cases with a Lag of One Month.
The comprehensive and summarized performance results for testing, and validation datasets, is displayed in Figures 2, and 3 respectively. These figures provide a visual picture of the model's performance across several levels of evaluation.

Table 2. Result Analysis

Feature Selection Method	Stopping Early at Epoch with minimum validation loss	RMSE Value			R2 Score		
		Training	Validation	Testing	Training	Validation	Testing
All Features	71 with validation loss 0.0011	0.02	0.03	0.06	0.87	0.86	0.78
RFE(Top 9)	73 with validation loss 0.0012	0.02	0.03	0.05	0.85	0.85	0.82
RFE(Top 8)	79 with validation loss 0.00125	0.02	0.03	0.05	0.85	0.86	0.85
RFE(Top 7)	89 with validation loss 0.00127	0.02	0.03	0.05	0.83	0.86	0.81
Lasso (Top 7)	89 with validation loss 0.00146	0.02	0.03	0.05	0.81	0.83	0.83
Lasso (Top 8)	79 with validation loss 0.00136	0.02	0.03	0.05	0.84	0.85	0.84
Correlation Coefficient (7)	89 with validation loss 0.00158	0.02	0.03	0.06	0.80	0.82	0.74
Correlation Coefficient (8)	62 with validation loss 0.00161	0.02	0.04	0.06	0.83	0.81	0.73

Journal of Information Systems Engineering and Management
2025, 10(34s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

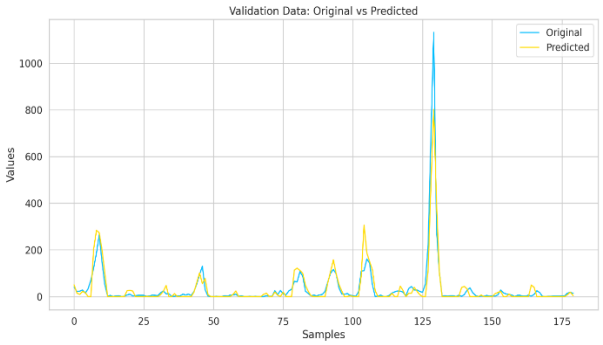

Figure 2: Predicted values for Validation dataset

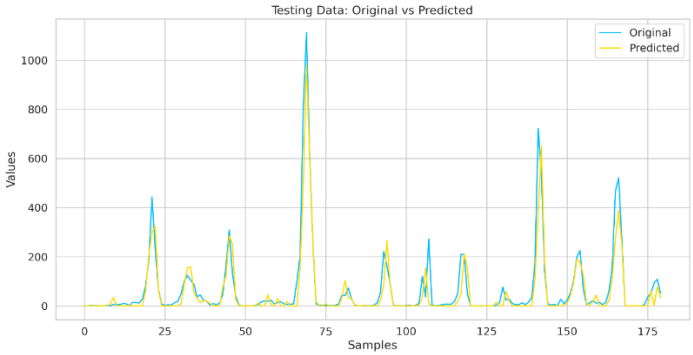

Figure 3: Predicted values for testing dataset

CONCLUSION

The study throws lights on the critical role of feature selection techniques in enhancing the predictive performance of LSTM models for dengue case forecasting. By employing Recursive Feature Elimination (RFE), we have identified key performance parameters or indicators which are maximum temperature, minimum temperature, average temperature, total monthly rainfall, total rainfall with a lag of one-month, average humidity, dengue cases with a lag of twelve months, and dengue cases with a lag of one month that significantly contribute to a higher R^2 score. The results indicates that these features, when integrated into the model, enables more accurate and reliable predictions. Additionally, the adoption of early stopping criteria based on the minimum validation loss, instead of a fixed number of epochs, effectively mitigates the risk of over fitting. This approach ensures that the model's performance generalizes well to unseen data and providing the robust results during testing.

REFRENCES

[1] G. Mussumeci, Elisa, and Flavio Codeco Coelho. (2020) “Large-scale multivariate forecasting models for Dengue - LSTM vs random forest regression.” Spatial and Spatio-temporal Epidemiology 35: 1-11. https://doi.org/10.1016/j.procs.2021.12.131

[2] Santosh, Thakur, Dharavath Ramesh, and Damodar Reddy. (2020) “LSTM based prediction of malaria abundances using big data.” Computers in Biology and Medicine 124: 1-8. https://doi.org/10.1016/j.compbiomed.2020.103859

[3] Ding, Yukai, Yuelong Zhu, Jun Feng, Pengcheng Zhang, and Zirun Cheng. (2020) “Interpretable spatio-temporal attention LSTM model for flood forecasting.” Neurocomputing 403: 348-359. https://doi.org/10.1016/j.neucom.2020.04.110

[4] Chang, Yue-Shan, Hsin.-Ta Chiao, Satheesh Abimannan, Yo-Ping Huang, Yi-Ting Tsai, and Kuan-Ming Lin. (2020) “An LSTMbased aggregated model for air pollution forecasting.” Atmospheric Pollution Research 11: 1451-1463. https://doi.org/10.1016/j.apr.2020.05.015

[5] Chandrashekar, Girish, and Ferat Sahin. "A survey on feature selection methods." Computers & electrical engineering 40.1 (2014): 16-28. https://doi.org/10.1016/j.compeleceng.2013.11.024

[6] Dewi, Christine, and Rung-Ching Chen. "Random forest and support vector machine on features selection for regression analysis." Int. J. Innov. Comput. Inf. Control 15.6 (2019): 2027-2037. 10.24507/ijicic.15.06.2027

[7] Bria, Yulianti Paula, et al. "Determining important features for dengue diagnosis using feature selection methods." medRxiv (2024): 2024-05. https://doi.org/10.1101/2024.05.05.24306901

[8] Dutta, Pijush, et al. "Feature selection based artificial intelligence techniques for the prediction of COVID like diseases." Journal of Physics: Conference Series. Vol. 1963. No. 1. IOP Publishing, 2021. 10.1088/1742-6596/1963/1/012167

[9] Hamada, Mohamed, et al. "Evaluation of Recursive Feature Elimination and LASSO Regularization-based optimized feature selection approaches for cervical cancer prediction." 2021 IEEE 14th International Symposium on Embedded Multicore/Many-core Systems-on-Chip (MCSoC). IEEE, 2021. https://doi.org/10.1109/MCSoC51149.2021.00056

Research Article

[10]Zeng, Longhui, and Zheng Chen. "Screening of genes characteristic of pancreatic cancer by LASSO regression combined with support vector machine and recursive feature elimination, and immune correlation analysis." Journal of International Medical Research 52.3 (2024): 03000605241233160. https://doi.org/10.1177/03000605241233160

[11] Fahmi, Amiq, et al. "Enhancing Prediction Accuracy in an Imbalanced Dataset of Dengue Infection Cases Using a Two-layer Ensemble Outlier Detection and Feature Selection Technique." International Journal of Intelligent Engineering and Systems 17.2 (2024): 544-560. 10.22266/ijies2024.0430.44

[12]Ge, Chenyu, et al. "FRL: An integrative feature selection algorithm based on the fisher score, recursive feature elimination, and logistic regression to identify potential genomic biomarkers." BioMed research international 2021.1 (2021): 4312850. https://doi.org/10.1155/2021/4312850

[13]Mehta, A. M., and K. S. Patel. "LSTM-based Forecasting of Dengue Cases in Gujarat: A Machine Learning Approach." Indian Journal of Science and Technology 17.7 (2024): 635-642. https://doi.org/10.17485/IJST/v17i7.2748

[14]Prechelt, Lutz. "Early stopping-but when?." Neural Networks: Tricks of the trade. Berlin, Heidelberg: Springer Berlin Heidelberg, 2002. 55-69. https://doi.org/10.1007/3-540-49430-8_3