Proceedings of 14th IOE Graduate Conference
Peer Reviewed
Year: 2023 Month: December Volume: 14
ISSN: 2350-8914 (Online), 2350-8906 (Print)

LSTM based Time Series Forecasting of Dengue

Suwarna Lingden [a], Anil Verma [b], Pukar Karki [c], Manoj Kumar Guragai [d]

[a, b, c, d] *Department of Electronics and Computer Engineering, Purwanchal Campus, IOE, Tribhuvan University, Nepal*

[a] suwarna_mise004@ioepc.edu.np, [b] anil@ioe.edu.np, [c] pukar@ioepc.edu.np, [d] manojkguragai@ioepc.edu.np

Abstract

Dengue fever, a mosquito-transmitted viral disease, presents a significant public health challenge in various regions, including Dharan, Nepal. Accurately predicting Dengue cases is vital for effective disease management and resource allocation. This research employs Long Short-Term Memory (LSTM) neural networks, utilizing meteorological and epidemiological data, to forecast Dengue cases in Dharan. LSTM models, known for capturing temporal dependencies, are designed, fine-tuned, and validated using various performance metrics.The research extends to real-time forecasting by incorporating the latest meteorological and epidemiological data. The LSTM-based model serves as a decision support tool for local public health authorities, aiding in scenario analysis and intervention planning. Noteworthy findings include RMSE = 8.061 and MAE = 6.507, which reflect the model's effectiveness.This research contributes to disease forecasting in resource-constrained settings and facilitates evidence-based interventions.

Keywords

LSTM, Dengue, Time Series Forecasting

1. Introduction

Nestled in the Sunsari district of south-eastern Nepal, Dharan stands as a captivating city characterized by a harmonious blend of cultural diversity, breathtaking natural beauty, and formiddable health challenges. Situated in the foothills of the Himalayas, Dharan boasts a subtropical climate with distinct wet and dry seasons. This climatic diversity plays a crucial role in the prevalence of vector-borne diseases such as Dengue[1].In recent times, Dharan has experienced an increasing danger from Dengue fever, which is a viral illness transmitted to humans by mosquitoes, primarily the Aedes mosquito vector. Dengue outbreaks have become recurrent in this region, causing a considerable public health burden, economic strain, and community distress.

Given its unique geographical location and climate, as well as the challenges posed by urbanization and vector-borne diseases, Dharan presents an ideal context for exploring innovative approaches to combat Dengue outbreaks. Leveraging meteorological data and advanced LSTM-based time series analysis holds the potential to provide valuable insights into the dynamics of Dengue transmission in Dharan, enabling better preparedness, response, and ultimately, improved public health outcomes in this dynamic and diverse city.

The motivation behind undertaking this research is driven by the urgent need to address the recurrent Dengue outbreaks in Dharan. These outbreaks have had a profound impact on public health in the region, resulting in a significant number of cases and hospitalizations during peak transmission periods. The gravity of Dengue's health implications, including severe illness and potential complications, underscores the pressing nature of finding effective preventive and control measures.

In addition to the immediate public health concerns, this research is motivated by the broader context of climate change. Dharan's susceptibility to Dengue is influenced by its climate, making it a potential hotspot for disease transmission. Understanding and addressing vector-borne diseases like Dengue are critical for future preparedness, given the changing climate patterns.

2. Literature Review

The key findings and research developments related to LSTM-based time series forecasting for Dengue outbreaks, with a particular focus on the unique context of Dharan, Nepal have been explored. It encompasses studies on Dengue epidemiology, the impact of meteorological factors, and the application of machine learning techniques for disease prediction.

2.1 Dengue Epidemiology and Impact in Nepal

Dengue fever, which is caused by the Dengue virus and mainly spread through the Aedes mosquito, has been increasingly prevalent in Nepal. While the disease was relatively rare in the past, the last decade has witnessed a notable increase in Dengue cases across the country [2]. A study highlights this growing concern, emphasizing the emergence of Dengue as a major public health issue in Nepal [3].

Furthermore, research conducted by Acharya et al. in 2019 has shed light on regional variations in Dengue transmission within Nepal. This study indicates that the disease's incidence patterns can differ significantly across regions, underscoring the importance of localized analysis and prediction strategies [4].

2.2 Meteorological Factors and Disease Transmission

Meteorological factors, including temperature and rainfall, play a crucial role in the transmission dynamics of Dengue fever. Studies such as the one conducted by Stewart-Ibarra et al. in 2018 have investigated the relationship between climate

and Dengue incidence. Their findings highlight how temperature and rainfall influence the abundance and behavior of Aedes mosquitoes, directly impacting Dengue transmission rates [5].

Additionally, research by Gharbi et al. in 2011 and Hii et al. in 2012 has delved into the seasonal patterns of Dengue outbreaks. These studies reveal the significance of temperature and rainfall in driving the seasonality of the disease, with variations in climate conditions influencing mosquito breeding and survival [6, 7].

2.3 Machine Learning and Disease Prediction

Recent advancements in machine learning have paved the way for the development of accurate predictive models for disease outbreaks.An effective method involves using Long Short-Term Memory (LSTM) neural networks for time series forecasting. LSTM models have proven to be successful in capturing intricate temporal patterns in data, making them well-suited for tasks related to predicting diseases.

Xu et al. in 2020 have successfully applied LSTM-based models to predict Dengue outbreaks in Singapore, showcasing the potential of these techniques in real-world disease forecasting scenarios [8].

S. Patil and S. Pandya in 2021 used time series models to forecast dengue hotspots associated with variation in meteorological parameters. Their discovery revealed that time-series forecasting models perform better than regression models like random forest regression and decision trees regression [9].

Majeed et. al. in 2023 used LSTM model and suggested temporal attention addition for LSTM models for dengue prediction. They claimed that both the LSTM and stacked LSTM (S-LSTM) models exhibited nearly identical performance. However, the addition of the attention mechanism significantly enhanced accuracy [10].

In 2020, U Khaira et. al. conducted a study comparing SARIMA and LSTM models for predicting the incidence of dengue hemorrhagic fever in Jambi, Indonesia. Based on the experiment both SARIMA and LSTM models were found to perform relatively well with LSTM performing slightly better than ARIMA [11].

Othman et. al. in 2022 developed a predictive model for dengue fever cases in Surabaya city through time series analysis. Their research revealed that the LSTM model effectively captured temporal patterns and accurately predicted future occurrences of the disease [12].

Doni Anjelus Ronald and Sasipraba Thankappan in 2020 used LSTM model for prediction of Dengue cases in India.The ReLU activation function was utilized, and upon training the LSTM model, it achieved an accuracy level of over 89% in forecasting infections during the epidemic and 81% for predicting deaths [13].

In 2020, Elisa Mussumeci and Flávio Codeço Coelho conducted extensive multivariate forecasting studies for Dengue, employing LSTM and random forest regression models. Among these methods, LSTM demonstrated superior performance in predicting future Dengue incidence in cities of varying sizes [14].

Alassafi et. al. in 2021 utilized LSTM for time series prediction of COVID-19. The LSTM models exhibited an accuracy of 98.58%, surpassing other RNN models that achieved only 93.45% accuracy [15].

Similarly,different studies on time series forecasting of diseases with the dataset of similar nature carried out by various researchers [16, 17, 18, 19, 20, 21] across the globe suggests the appropriateness and effectiveness of using this model for the forecasting of dengue in Dharan.

3. Methodology

3.1 Data Collection

Both the meteorological and epidemiological data were obtained from secondary references, and these sources will be detailed in the following sections.

3.1.1 Meteorological Data

Historical meteorological data were obtained from the Office of Hydrology and Meteorology Science,Eastern Regional Climate Office,Dharan,Nepal.This data included precipitation,maximum temperature,minimum temperature,and relative humidity recorded on a daily basis for the last five years. The time series plot of these meteorological features are depicted in Figure 1,Figure 2,Figure 3,and Figure 4 respectively for the date range of April 2019 to September 2023.

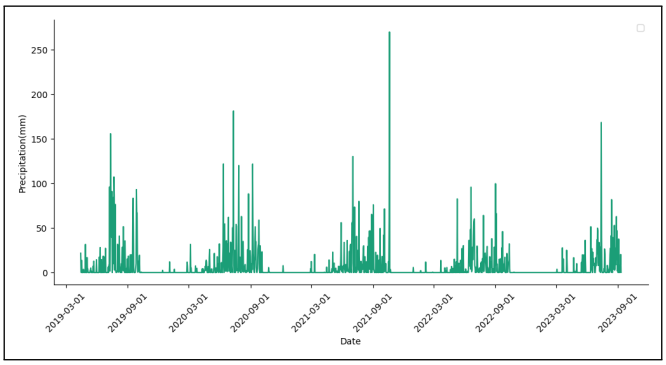

Figure 1: Time series plot of precipitation in mm

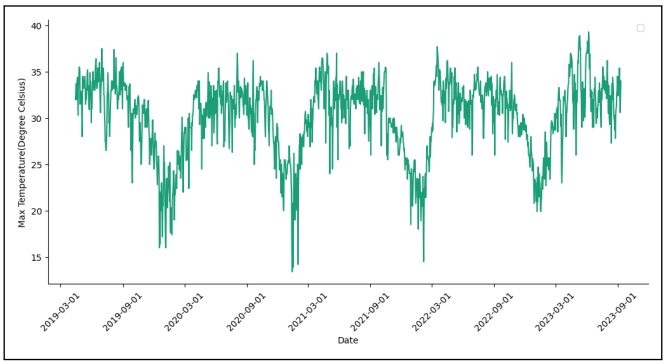

Figure 2: Time series plot of max temperature in degree celsius

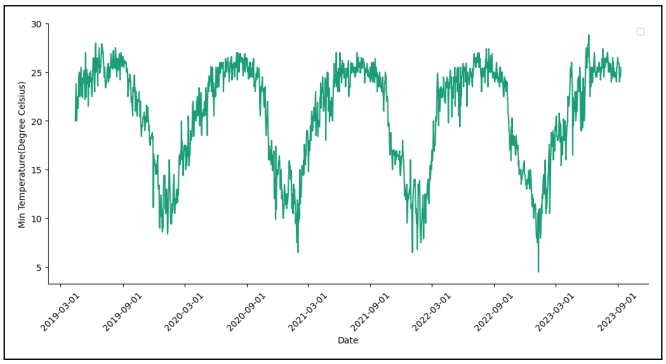

Figure 3: Time series plot of min temperature in degree celsius

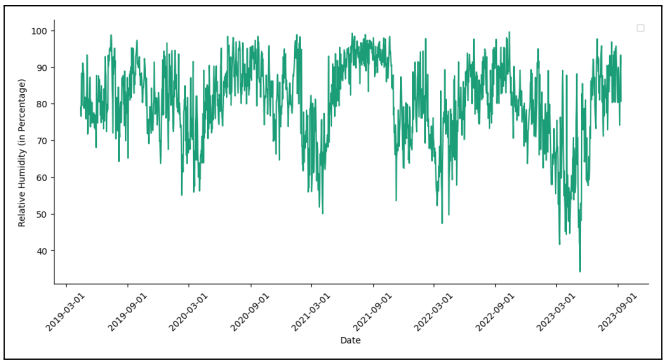

Figure 4: Time series plot of relative humidity in percentage

3.1.2 Epidemiological Data

Historical Dengue incidence records including the dates of occurrence and the number of reported cases for Dharan were maintained by the Health Division of Dharan Sub-Metropolitan City Municipal Office.

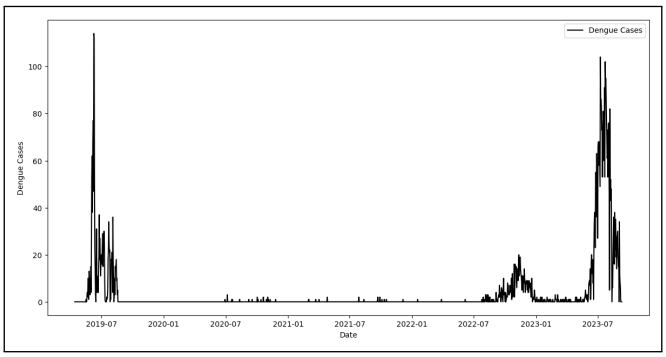

Figure 5: Time series plot of daily dengue cases from 2019 to 2023

3.2 Data Pre-processing

3.2.1 Data cleaning

There were no any missing values in the selected date range.Hence,efforts were not needed to handle any missing values. However,there were outliers which have been adjusted with the help of the domain knowledge.

3.2.2 Data Transformation

Feature scaling was done with the help of MinMaxScaler to normalize numerical features to have similar scales.The date object in the dataset was changed to date and time format.

3.2.3 Windowing

An overlapping window of seven historical data were considered.This window acted as an input sequence for the LSTM model,allowing it to learn patterns and dependencies in the data.

3.2.4 Correlation Matrix

The correlation matrix of meteorological features and Dengue cases is as shown in the Figure 6.

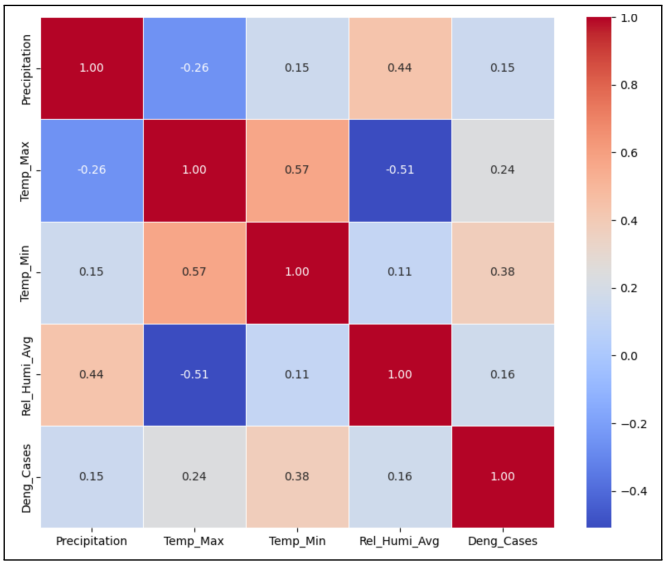

Figure 6: Correlation matrix of meteorological features and Dengue Cases

The relationships between weather variables and dengue cases were analyzed, revealing varying degrees of correlation. Precipitation levels and maximum temperature showed weak positive correlations (0.15 and 0.24, respectively), indicating a slight association with higher dengue cases. The relationship with relative humidity also demonstrated a weak positive correlation (0.16). However, minimum temperature exhibited a moderate positive correlation (0.38), suggesting a stronger association with increased dengue cases compared to the other weather variables. Overall, while these weather factors are linked to dengue cases, the associations are generally weak, with minimum temperature showing a comparatively stronger correlation.

3.3 Building the training data

Two empty lists were created and were populated with training data.One stored the input sequences (features), and the other stored the corresponding target values.The number of future values to predict was set to one.The data to be used as input features from the past was set to seven.A loop function was designed that iterated through the dataset starting from the past seventh day to the end of the dataset

minus the first day to predict.This loop function was used to create training sequences. Finally the sequences were converted into arrays for further processing.

3.4 Building the LSTM Model

A sequential model which is a linear stack of layers where one layer can be added at a time was used.The first layer to the model,an LSTM layer with 32 memory units was designed.The ReLU (Rectified Linear Unit) activation function was used in this layer.

A dropout layer was added after the LSTM layer. Dropout is a regularization technique used to prevent overfitting. It randomly sets a fraction (in this case, 0.2) of the input units to zero during training, which helped improve the model's generalization.

The Adam optimizer was selected as the optimization algorithm, and Mean Squared Error (MSE) was used as the loss function. MSE calculates the average squared disparity between the predicted values and the actual target values.

In summary,a sequential neural network model with an LSTM layer for time series forecasting was designed that used ReLU activation, dropout regularization, and MSE loss for training. The model was designed to take sequences of past data as input and predict a single value (Dengue cases) for the next day.

3.5 Training the LSTM Model

An epoch of 50 was set. Batch size of 16 was chosen.10 percent of training data was used for the validation purpose.The experiment focussed on forecasting dengue cases using historical meteorological and epidemiological data. The LSTM model was trained using the prepared dataset, with dengue cases as the target variable.

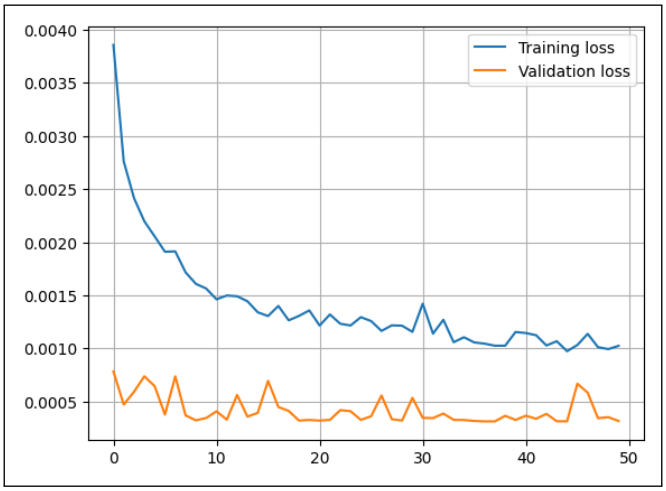

Figure 7: Training loss vs Validation loss for 50 epochs

4. Model Evaluation

Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) are frequently employed metrics for assessing the performance of models in regression tasks. These metrics offer valuable information regarding how accurately a predictive model can estimate or predict continuous numeric values.

4.1 Mean Absolute Error(MAE)

The Mean Absolute Error (MAE) is a statistical metric that calculates the average magnitude of the differences between the actual values and the predicted values. Mathematically, it is represented as follows.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{1}$$

n = number of data points or observations

Y_i = actual values (ground truth) for the i-th data point

$\hat{y}_i$ = predicted values for the i-th data point

4.2 Root Mean Square Error(RMSE)

Root Mean Square Error (RMSE) is a commonly used metric in statistical analysis and machine learning to measure the accuracy of a predictive model.Its mathematical representation is as shown below:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{2}$$

n = The total number of data points or observations.

y_i = The actual (observed) value for the i-th data point.

$\hat{y}_i$ = The predicted value for the i-th data point.

5. Results and Discussion

The LSTM model successfully predicted Dengue cases in Dharan from 2019 to 2023, achieving a Mean Absolute Error (MAE) of 6.507 and a Root Mean Squared Error (RMSE) of 8.061. These metrics indicate that, on average, the model's predictions deviated by approximately 6.5 Dengue cases, and the deviations were, on average, within an 8-case range from the actual values. The model's performance was notable, especially given the complexity of Dengue transmission dynamics.

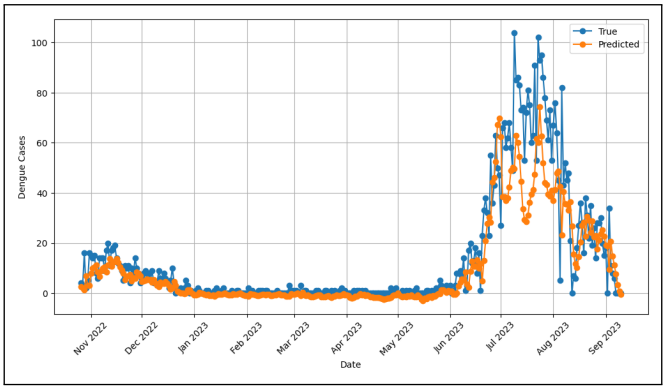

Figure 8: Time series forecasting of Dengue Cases for the test set date range

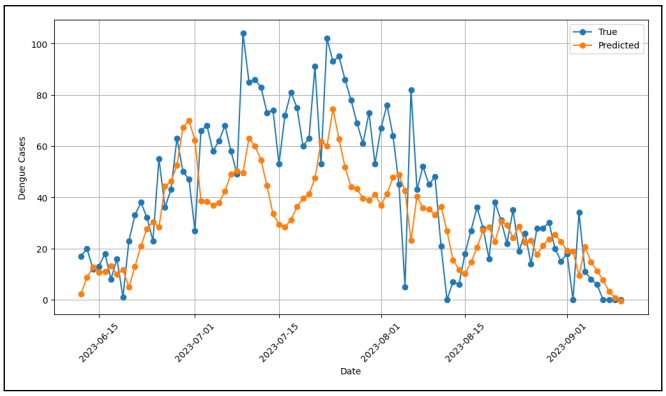

Figure 9: Forecasting of the Dengue Cases of last 90 days before September 10, 2023

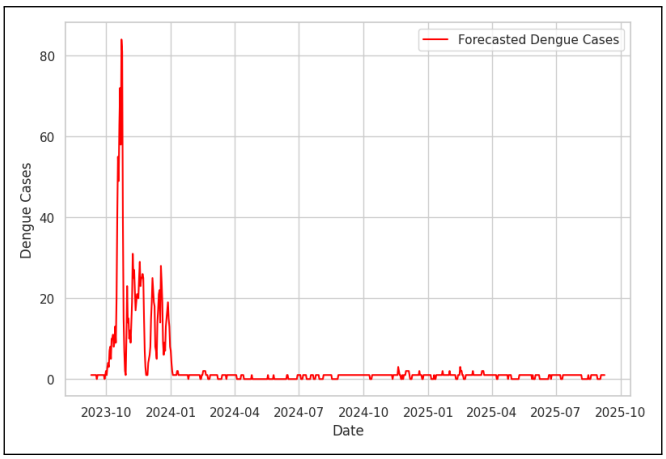

Figure 10: Future Forecast Upto September 2025

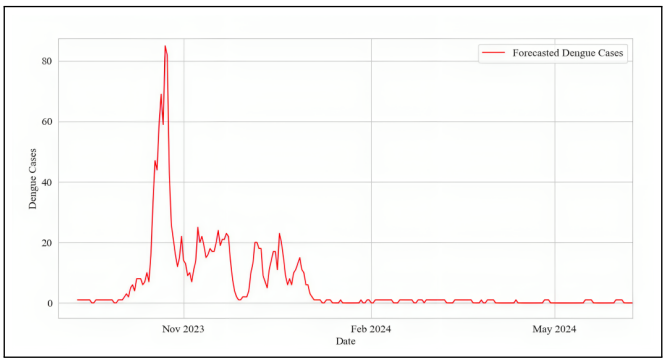

Figure 11: Future Forecast from September 2023 to June 2024

A visual comparison between the actual and predicted Dengue cases as depicted in Figure 8 and Figure 9 revealed a high degree of alignment in most periods. However, during certain months, particularly those coinciding with the monsoon season, the model exhibited discrepancies between predicted and actual cases. These deviations are crucial for understanding the model's limitations and potential areas of improvement.

Upon closer inspection, outliers in the prediction errors were observed during specific periods of heavy rainfall. The model tended to overestimate Dengue cases during these rainy months, indicating a potential sensitivity to extreme weather events.

The model demonstrated a tendency to underestimate cases during periods of relative stability in meteorological conditions. This underestimation might be attributed to the model's difficulty in capturing subtle changes in environmental factors when disease transmission remains relatively constant.

Figure 10 and Figure 11 shows the forecast for the future dates.The forecasting of future dengue cases for the next two years based on historical meteorological data from the preceding two years was a focal point. This approach, relying on past values, provided a foundation for anticipating the potential trajectory of dengue outbreaks. By employing a Long Short-Term Memory (LSTM) model, the analysis leveraged the temporal patterns in the dataset, offering a reliable tool for projecting future scenarios. The consideration of meteorological factors, such as precipitation, temperature, and humidity, in the forecasting process strengthened the model's predictive capabilities. This strategy of forecasting future dengue cases based on recent historical data emerges as a valuable approach for proactive public health planning and intervention strategies.

The selected meteorological features were chosen based on their established impact on Aedes mosquito abundance and Dengue transmission. While these features proved valuable, further exploration of additional environmental variables, such as vegetation indices and water stagnation data, could enhance the model's accuracy by capturing finer ecological nuances.

6. Conclusion

In this study, the LSTM-based predictive model has demonstrated promising results, achieving a low Mean Absolute Error (MAE) of 6.507 and a Root Mean Squared Error (RMSE) of 8.061. These metrics indicate the model's capacity to generate accurate forecasts for dengue case counts, which holds significant value for public health authorities and healthcare organizations.However, it is essential to acknowledge the limitations associated with this study, primarily the dataset's size. Despite these challenges, this study underscores the potential of dengue prediction models in public health.In conclusion, this study has made significant progress in predicting and mitigating dengue outbreaks.

7. Future enhancements

Future enhancements for this research endeavour encompass three vital aspects: dataset expansion, broader predictive capabilities, and model refinement. Firstly, the study will benefit from a more extensive and diverse dataset that incorporates data from various regions, dengue strains, and a longer time frame. This larger dataset will enable the model to capture a more comprehensive range of factors affecting dengue dynamics and improve overall accuracy.Secondly, the study will extend its predictive capabilities beyond simple case count forecasting. It will include dengue outbreak prediction, aiming to provide early warnings of potential outbreaks by detecting early trends and anomalies in the data. Additionally, the study will incorporate severity prediction,

assessing the intensity and impact of outbreaks.Lastly, model refinement remains a crucial focus. This includes fine-tuning the LSTM-based model for optimal performance, exploring ensemble models that combine multiple forecasting algorithms, and integrating external factors population density,and vaccination rates. Real-time data integration will ensure that the model remains current and relevant, aiding timely decision-making.

Acknowledgments

This study is carried out under the Department of Electronics and Computer Engineering at IOE, Purwanchal Campus, Dharan. The authors wish to extend their sincere appreciation to all the faculty members in the department for their invaluable contributions to this endeavor.

References

[1] Dharan Sub-Metropolitan Municipal office. https://www.dharan.gov.np/ne/node/14. 2023.

[2] Epidemiology and Disease Control Division. https://www.edcd.gov.np/section/dengue-control-program. 2023.

[3] Shyam Prakash Dumre, Dhiraj Acharya, Bibek Kumar Lal, and Oliver J Brady. *Dengue climbing on top of the world-2019's big jump broke the record.* 2019.

[4] Bipin Kumar Acharya, W. Chen, and Z. Ruan. *Mapping environmental suitability of scrub in Nepal using MaxEnt and Random Forest models.* 2019.

[5] Rachel Lowe, Antonio Gasparrini, Cedric J. Van Meerbeeck, Catherine A. Lippi, Roche Mahon, Adrian R. Trotman, Leslie Rollock, Avery Q. J. Hinds, Sadie J. Ryan, and Anna M. Stewart-Ibarra. *Nonlinear and delayed impacts of climate on dengue risk in Barbados: A modelling study.* 2018.

[6] Myriam Gharbi, Philippe Quenel, Joël Gustave, Sylvie Cassadou, Guy L. Ruche, Laurent Girdary, and Laurence Marrama. *Time series analysis of dengue incidence in guadeloupe, french west indies: Forecasting models using climate variables as predictors.* 2011.

[7] Yien Ling Hii, Huaiping Zhu, Nawi Ng, Lee Ching Ng, and Joaciam Rocklov. *Forecast of Dengue Incidence Using Temperature and Rainfall.* 2012.

[8] Jiucheng Xu, Keqiang Xu, Zhichao Li, Fengxia Meng, Taotian Tu, and Qiyong Liu. *Forecast of Dengue Cases in 20 Chinese Cities Based on the Deep Learning Method.* 2020.

[9] S Patil and S Pandya. *Front. Public Health*, 9:798034, 2021.

[10] Mokhalad A Majeed, Helmi ZM Shafri, Aimrun Wayayok, and Zed Zulkafli. *Prediction of dengue cases using the attention-based long short-term memory (LSTM) approach. Geospatial health*, 18(1), 2023.

[11] Ulfa Khaira, Pradita Eko Prasetyo Utomo, Reni Aryani, and Indra Weni. *A comparison of SARIMA and LSTM in forecasting dengue hemorrhagic fever incidence in Jambi, Indonesia.* In *Journal of Physics: Conference Series*, volume 1566, page 012054. IOP Publishing, 2020.

[12] Mahmod Othman, Rachmah Indawati, Ahmad Abubakar Suleiman, Mochammad Bagus Qomaruddin, and Rajalingam Sokkalingam. *Model Forecasting Development for Dengue Fever Incidence in Surabaya City Using Time Series Analysis. Processes*, 10(11):2454, 2022.

[13] Anjelus Ronald Doni and Thankappan Sasipraba. *LSTM-RNN Based Approach for Prediction of Dengue Cases in India. Ingénierie des Systèmes d'Information*, 25(3), 2020.

[14] Elisa Mussumeci and Flávio Codeço Coelho. *Large-scale multivariate forecasting models for Dengue-LSTM versus random forest regression. Spatial and Spatio-temporal Epidemiology*, 35:100372, 2020.

[15] Madini O Alassafi, Mutasem Jarrah, and Reem Alotaibi. *Time series predicting of COVID-19 based on deep learning. Neurocomputing*, 468:335–344, 2022.

[16] Zulfany Erlisa Rasjid, Reina Setiawan, and Andy Effendi. *A comparison: prediction of death and infected COVID-19 cases in Indonesia using time series smoothing and LSTM neural network. Procedia computer science*, 179:982–988, 2021.

[17] Refat Khan Pathan, Munmun Biswas, and Mayeen Uddin Khandaker. *Time series prediction of COVID-19 by mutation rate analysis using recurrent neural network-based LSTM model. Chaos, Solitons & Fractals*, 138:110018, 2020.

[18] Mohamed Djerioui, Youcef Brik, Mohamed Ladjal, and Bilal Attallah. *Heart Disease prediction using MLP and LSTM models.* In *2020 International Conference on Electrical Engineering (ICEE)*, pages 1–5. IEEE, 2020.

[19] Peipei Wang, Xinqi Zheng, Gang Ai, Dongya Liu, and Bangren Zhu. *Time series prediction for the epidemic trends of COVID-19 using the improved LSTM deep learning method: Case studies in Russia, Peru and Iran. Chaos, Solitons & Fractals*, 140:110214, 2020.

[20] Yu-Tse Tsan, Der-Yuan Chen, Po-Yu Liu, Endah Kristiani, Kieu Lan Phuong Nguyen, and Chao-Tung Yang. *The prediction of influenza-like illness and respiratory disease using LSTM and ARIMA. International Journal of Environmental Research and Public Health*, 19(3):1858, 2022.

[21] Vinay Kumar Reddy Chimmula and Lei Zhang. *Time series forecasting of COVID-19 transmission in Canada using LSTM networks. Chaos, solitons & fractals*, 135:109864, 2020.