ENHANCING DENGUE FEVER PREDICTIONS IN TROPICAL CITIES: A COMPARATIVE ANALYSIS OF MACHINE LEARNING MODELS IN SAN JUAN AND IQUITOS

J. Jiang, W. Sun, Z. An

University of Nottingham
School of Computer Science
Nottingham, United Kingdom

ABSTRACT

This study explores the efficacy of machine learning (ML) techniques for forecasting dengue prevalence, utilizing data from San Juan and Iquitos. A thorough evaluation was conducted to ascertain the most adept model for weekly dengue case predictions, applying a suite of regression models: Linear Regression(LR), Random Forest(RF), Multi-layer Perceptron(MLP), Long Short-Term Memory (LSTM) networks, and XGBoost. The critical role of data preprocessing and feature selection in enhancing model precision is underscored. The findings reveal that the XGBoost and LSTM models, augmented by feature selection methods like recursive feature elimination, markedly surpass other models in terms of mean absolute error (MAE). These results underscore their viability in refining dengue monitoring and bolstering public health measures.

***Index Terms*—** Dengue Fever, LSTM, XGBoost, Machine Learning, Prediction

1. INTRODUCTION

Dengue fever is a group of vector-borne infectious diseases caused by the dengue virus (DV), prevalent in tropical and subtropical regions. Over the recent decades, there has been a significant increase in global incidence. The World Health Organization estimates that there are about 50 million DV infections worldwide each year, with 500,000 people requiring hospitalization for dengue fever, and approximately 2.5% of infections resulting in death [1]. Currently, dengue fever is endemic in over 100 countries across Africa, the Americas, the Eastern Mediterranean, Southeast Asia, and the Western Pacific.

Predicting infectious epidemics such as dengue fever is a challenging task. Some prediction technologies are still in their early stages, but machine learning algorithm research is gradually receiving academic attention and research. Valuable trend information can be extracted from the data through standardized data sets, thereby enabling the ability to systematically predict dengue fever outbreaks [2]. In the presence of large-scale time-series data, machine learning techniques are employed to model the underlying patterns and extract meaningful insights. Leveraging the strengths of various algorithms, such as XGBoost, LSTM, LR, and RF and MLP, this study delves into the predictive capabilities of machine learning models for diverse applications.

With LR and RF serving as baseline models, the study demonstrates their performance benchmarks. For two cities with distinct geographical environments, the XGBoost regression model achieves an outstanding prediction accuracy of 13.42 (MAE) for the city of San Juan. Meanwhile, the LSTM regression model surpasses this performance in Iquitos, achieving a score of 5.93.

2. LITERATURE REVIEW

After analyzing the bibliometric map of 274 scientific articles and narrowing them down to 33 for further analysis, it was concluded that neural networks are the most frequently used in predictive models for detecting dengue fever infections [3]. In the study [4], a deep neural network predictive model was adopted and compared with traditional decision tree and logistic regression models, demonstrating the superiority of neural network models. With only four input variables, it achieved an area under the ROC curve (AUC) of 85.87%. It is evident that classical models, including neural network models, can further enhance their performance in dengue fever prediction.

'They say if things don't go your way in predictive modeling, use XGBoost. XGBoost, Extreme Gradient Boosting is an ensemble tree method developed by Tianqi Chen and Carlos Guestrin that applies the principle of boosting weak learners using the gradient descent architecture and is designed to focus on speed and efficiency. It works best for medium structured tabular data. [5]'. The XGBoost model performed well in Dengue prediction. This model, an ensemble tree method, focuses on boosting weak learners using a gradient descent architecture and is designed for speed and efficiency, particularly suited for medium structured tabular data. The study involved data processing, model building, tuning parameters, cross-validation, and calculation of met-

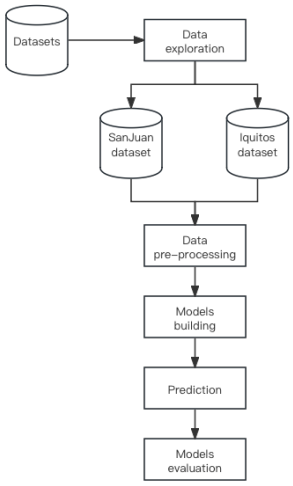

Fig. 1: Project Architecture

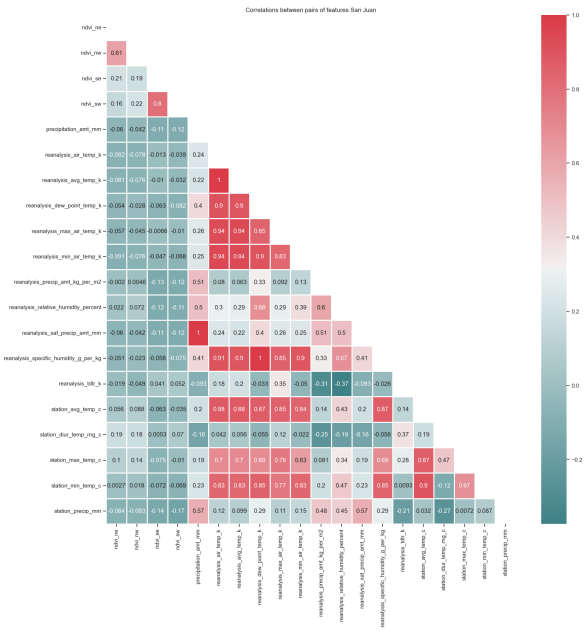

Fig. 2: Features Correlation Matrix of SanJuan

rics such as mean absolute error, median squared error, and root mean squared error for the XGBoost model. Therefore, the XGBoost model demonstrated good predictive capabilities in Dengue prediction. Additionally, achieving commendable results, with MAE ranging between 8 to 15, was also a considerable accomplishment.

The study [6] presents an innovative approach for forecasting dengue fever incidence in Malaysia by employing a Long Short-Term Memory (LSTM) network, a specialized form of recurrent neural network designed for deep learning tasks. The authors built an LSTM-based predictive model, utilizing a dataset which consist of meteorological and climatic variables, and benchmarked its performance against a conventional Support Vector Regression (SVR) model. The results indicate a superior predictive capability of the LSTM model, evidenced by an R2 score of 0.75 and a MAE of 8.76. This research corroborates the potential of LSTM models as robust analytical tools in enhancing dengue fever surveillance and aiding public health interventions strategies.

3. METHODOLOGY

This study formulates the problem as a regression task. To address this challenge, this research developed a pipeline that integrates various data preprocessing methods and applied it to multiple advanced regression learners, including Linear Regression (LR), Random Forest (RF), Long Short-Term Memory (LSTM) networks, Multi-layer Perceptron(MLP) and XGBoost. The overall architecture of the research methodology is presented in Fig. 1

3.1. Data Analysis and Pre-processing

3.1.1. Data Analysis

Initially, to account for the geographical disparities, the factors influencing dengue fever cases might differ between San Juan and Iquitos, therefore, the dataset was split to facilitate a detailed analysis of data categorized by city, allowing for individualized examination of each city's unique characteristics and patterns.

Descriptive statistical analysis was performed to calculate the mean, median, and mode for each numerical feature, providing an initial understanding of the data's central tendencies and distributions. For visual analysis, violin plots were employed to examine the distribution and range of each feature, while box plots were utilized to detect and identify outliers, a crucial step for the subsequent data cleaning process. In terms of statistical testing, the Shapiro-Wilk test was conducted to assess normality, thereby determining whether the data approximated a normal distribution. Additionally, the Pearson correlation coefficient was calculated to analyze the linear relationships between numerical variables, laying the groundwork for the ensuing feature selection process. They were shown in Fig. 2 and Fig. 3.

Furthermore, time series analysis was applied to dissect the major components such as trends, periodicity, and seasonality of the time series data. Finally, seasonal changes were meticulously identified and quantified to understand cyclical fluctuations and to aid in the prediction of future trends, enhancing the robustness of analytical approach.

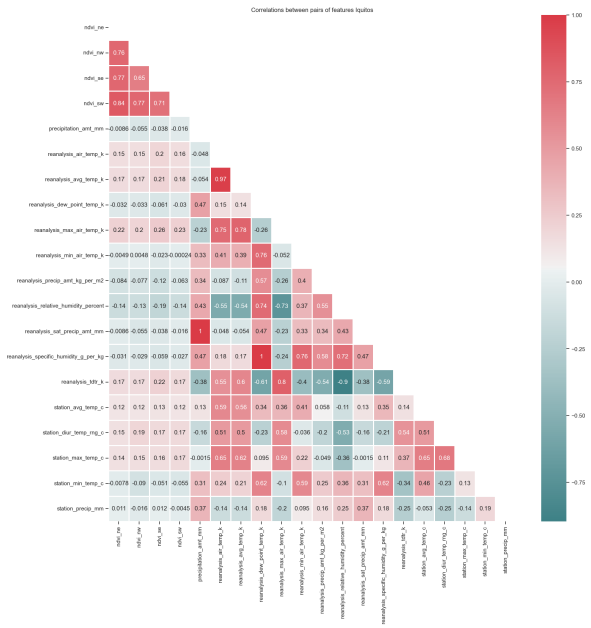

Fig. 3: Features Correlation Matrix of Iquitos

3.1.2. Data Pre-processing

Data pre-processing was undertaken to ensure the integrity and quality of the dataset [7]. It is worth noting that there are missing values in every feature, which poses a challenge for this time series problem. An appropriate method needs to be chosen to handle the missing values. Common methods include deleting missing values and filling missing values. Given that the efficacy of the subsequent analysis is contingent upon the refinement of raw data, meticulous attention was directed towards mistaken and missing data process. The forward fill was engaged for nan values, thereby ensuring the optimal functionality of the prediction algorithm.

Subsequently, the dataset was normalized using both MinMaxScaler, StandardScaler and RobustScaler in this research. The MinMaxScaler adjusts values to range from 0 to 1, ensuring equal treatment of all features, while the RobustScaler effectively handles outliers by scaling data based on the median and the interquartile range, thereby enhancing the robustness of the scaling process.

3.1.3. Feature Selection

In this study, multiple feature selection techniques were implemented to optimize predictive models for dengue fever case prediction. It utilized the Pearson correlation coefficient to eliminate features with minimal correlation to the target, thereby simplifying the dataset. Additionally, Decision Tree was employed to visually assess the impact of different features on predictions, enhancing understanding of their relationships and significance.

LASSO regression was also applied, a method that penalizes less important features, reducing them to zero to streamline and improve the model. Concurrently, Principal Component Analysis (PCA) was used to transform a large set of correlated variables into a smaller set of uncorrelated variables, capturing essential information while reducing complexity. Furthermore, Recursive Feature Elimination (RFE) was implemented, which iteratively builds models and removes the least significant feature each time until the most effective features are identified.

Gradient Boosting models, such as XGBoost and LightGBM, offer assessments of feature importance, which are generally based on the role features play during splits (such as the number of splits and the gains following the splits).

These techniques were thoroughly evaluated to determine their impact on the model's accuracy, sensitivity, and specificity. The integrated use of these methods significantly refined models, enhancing the capability to accurately predict dengue fever cases.

3.2. Modelling

3.2.1. Machine Learning Regression Techniques

1) Linear Regression

Linear regression (LR) [8] is a sophisticated statistical method that delineates the dynamics between a dependent variable and multiple independent variables through the application of a linear equation to the observed data. It assigns a unique coefficient to each independent variable, thereby quantifying its specific contribution to the dependent variable while all other variables remain constant.

2) Random Forest

Random forests consist of multiple tree predictors, where each tree is influenced by a randomly sampled vector. This vector is sampled independently and follows the same distribution across all trees within the forest [9].

3) Multilayer Perceptron

Multilayer Perceptron (MLP) is a fundamental form of feedforward artificial neural network comprising an input layer, multiple hidden layers, and an output layer. MLPs are versatile in addressing both classification and regression challenges, making them suitable for tasks like forecasting dengue fever outbreaks.

4) XGBoost

XGBoost is a powerful framework that employs the extreme gradient boosting algorithm, a method involving an ensemble of gradient-boosted decision trees. This approach iteratively enhances the model by sequentially integrating numerous weak classifiers to reduce the loss function. Favored for its efficiency, XGBoost leverages parallel processing and distributed computing, optimizing both computation time and resource usage.

Furthermore, XGBoost utilizes an ensemble learning approach to combine the predictions of multiple weak learners

(decision trees) to generate a robust and accurate model. This ensemble strategy helps mitigate the bias and variance of individual trees, resulting in more reliable predictions, which is particularly beneficial in complex prediction tasks such as disease transmission modeling. [10]

XGBoost effectively prevents model overfitting through regularization techniques, such as L1 and L2 regularization, thereby ensuring the model's ability to generalize to unseen data, which is crucial for accurately predicting disease transmission.

5) Long Short-term Memory (LSTM)

Given that the data were collected on a weekly basis, time series analysis techniques were deemed appropriate for use. LSTM are capable of capturing temporal trends and seasonality within the dataset, which could significantly enhance the accuracy of predictions. Long Short-Term Memory (LSTM) networks, a specialized variant of Recurrent Neural Networks (RNN), are engineered to overcome the limitations commonly associated with traditional RNNs, notably their struggle with learning long-term dependencies [11]. At the heart of LSTM networks is their ability to retain information for extended periods, making them highly effective for tasks involving sequential data, such as natural language processing and time series analysis.

3.2.2. Evaluation Metrics

Evaluation metrics are the key to measuring the performance of machine learning models. This experiment used MAE as the evaluation metric for the regression model.

Mean Absolute Error (MAE) is a metric used to evaluate the performance of regression models. MAE measures the average of the sum of the absolute differences between observation values and predicted values. The MAE calculation equation is as follows:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_{\text{true},i} - y_{\text{pred},i}| \tag{1}$$

4. RESULTS

This section showcases the results of this thorough investigation, emphasizing data analysis, preprocessing, and feature selection, alongside the efficacy of these regression models. It evaluated diverse methodologies proposed by different team members, with the goal of pinpointing the most efficient tactics for forecasting weekly dengue fever incidences in San Juan and Iquitos.

4.1. Data Analysis Results

The initial analysis involved examining the dataset comprising 1456 samples across 24 features. Among the 24 features, their sources are not the same. The dataset mainly consist of measurement data from NOAA's GHCN weather stations, PERSIANN satellite precipitation measurement data, NOAA's NCEP Climate Forecast System Reanalysis measurements, and satellite vegetation - Normalized Difference Vegetation Index (NDVI) in four parts.

Descriptive statistical analyses were conducted for these two cities, revealing insights into climate and vegetation indices over several years. For Iquitos, vegetation indices such as NDVI exhibited mean values around 0.26, and precipitation averaged 64.25 mm. Temperature-related metrics, like reanalysis air temperature, averaged 297.87 K. San Juan displayed a different climate profile, with lower precipitation averaging 35.47 mm and slightly higher temperatures, mean reanalysis air temperature being 299.16 K. These statistics provide foundational knowledge for understanding environmental patterns affecting dengue fever incidences in these regions.

The Shapiro Wilk test revealed significant deviations from normality in the precipitation related variables (precipitation_amt_mm, reanalysis_precip_amt_kg_per_m2, and reanalysis_sat_precip_amt_mm) for both cities. These deviations, characterized by heavy tails or outliers, are commonly associated with rainfall data. Additionally, the total case counts in both cities are markedly non-normal, possibly indicating the occurrence of disease outbreaks and epidemics that cause peaks in data distribution, deviating from typical patterns.

In San Juan, the variables reanalysis_relative_humidity_percent and station_diur_temp_rng_c displayed statistics suggesting they are relatively closer to a normal distribution compared to other features. In Iquitos, the variables reanalysis_air_temp_k and reanalysis_avg_temp_k demonstrated p-values that indicate a better adherence to normal distribution, potentially simplifying modeling and analytical efforts for these specific attributes.

In Fig. 4 and Fig. 5, violin plots were presented distributions for all features of two cities. Upon comparison, the four features of San Juan City: Precipitation_amt_mm, Reanalysis_precip_amt_kg_per_m2, Reanalysis_sat_precip_amt_mm, and Station_precip_mm, exhibit a wider spread of discrete points. This suggests that the city experiences a greater range of fluctuation in the target variable. In contrast, the overall distribution of Iquitos City appears more visually centered towards a normal distribution. The actual correlation will be determined in other statistical analyses.

The time series analysis of environmental and epidemiological data from San Juan and Iquitos revealed significant correlations between dengue cases and climatic variables. Seasonal spikes in temperature, precipitation, and humidity in both cities were closely aligned with increases in dengue incidence, indicating a strong environmental influence on disease transmission. Analysis also showed that vegetation indices, while less variable, provided essential context for understanding mosquito-borne disease dynamics. These insights highlight the importance of environmental monitoring

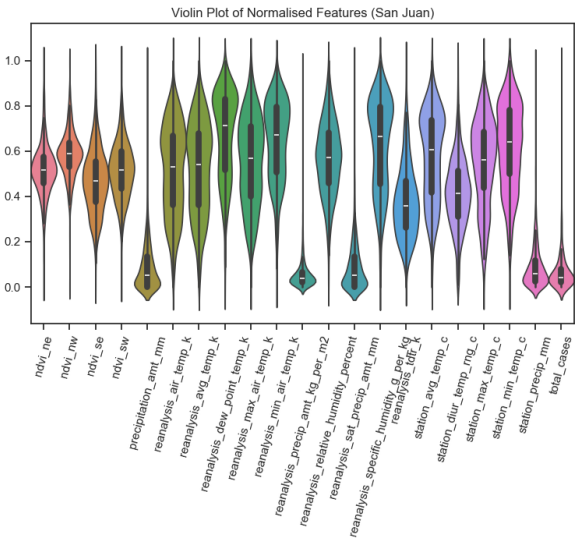

Fig. 4: Violin Plot of San Juan

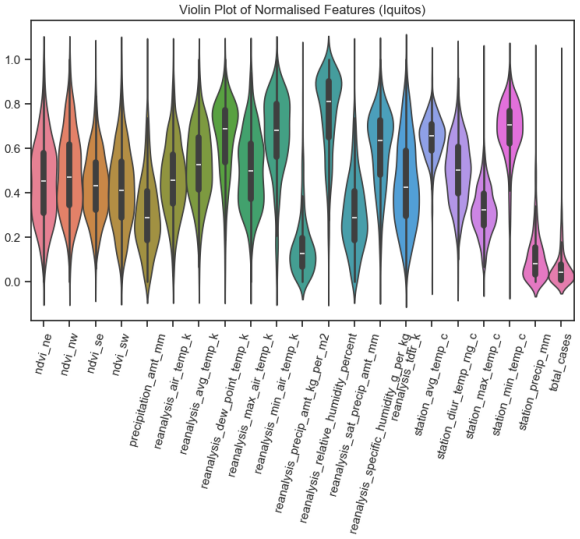

Fig. 5: Violin Plot of Iquitos

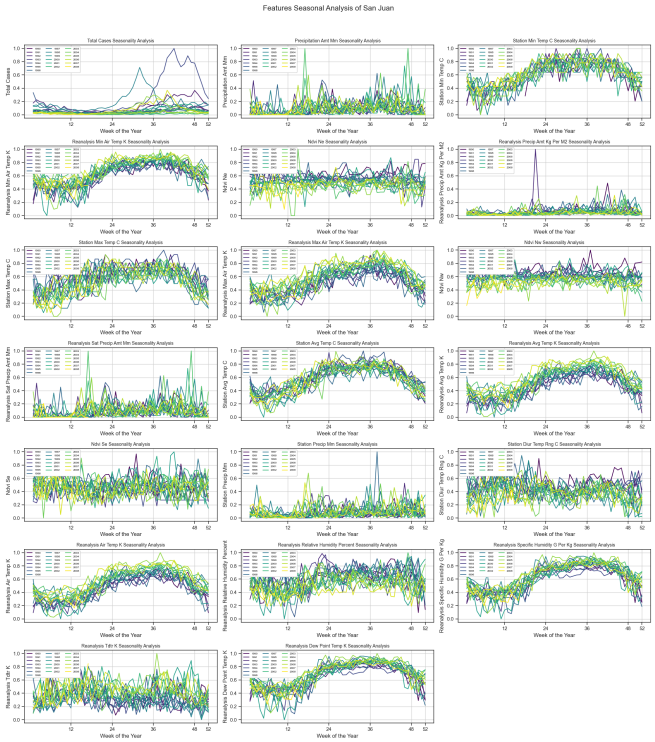

Fig. 6: Seasonality Analysis of San Juan

in predicting and managing dengue outbreaks, emphasizing the value of time series analysis in public health strategy development.

Seasonality analysis of climatic variables from San Juan (Fig. 6) and Iquitos (Fig. 7) reveals distinct patterns that align with fluctuations in dengue cases. In both cities, total dengue cases how significant seasonal fluctuations, which correlate strongly with specific environmental conditions. Notably, precipitation and temperature show clear seasonal cycles, suggesting their influence on dengue transmission dynamics. San Juan experiences more pronounced seasonal variations in temperature, whereas Iquitos shows greater variability in precipitation patterns. This analysis underscores the importance of understanding seasonal environmental changes to enhance predictive models for dengue outbreaks and inform public health strategies effectively.

4.2. Pre-processing and Feature Selection Results

The preprocessing stage plays a key role in improving model performance. In this study, erroneous week data were dealt with first, and the key techniques employed encompassed forward filling for managing missing data, along with MinMaxScaler and RobustScaler for data normalization.

For feature selection, a combination of strategies was utilized, including the Filter Method, LASSO, Decision Trees, Pearson Correlation Coefficient, and Wrapper Methods (specifically Recursive Feature Elimination, RFE). This

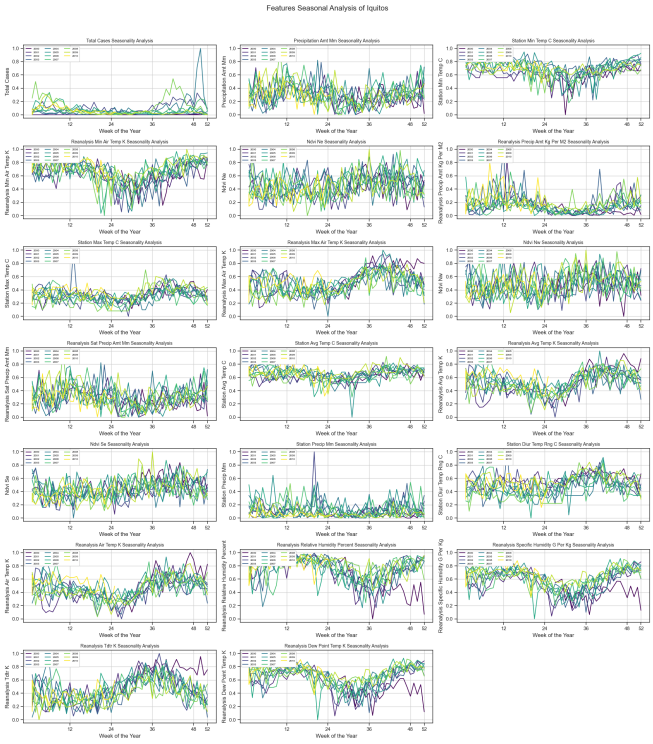

Fig. 7: Seasonality Analysis of Iquitos

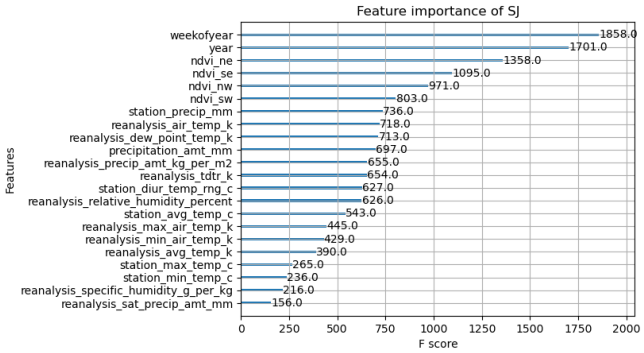

Fig. 8: Xgboost Feature Important of SanJuan

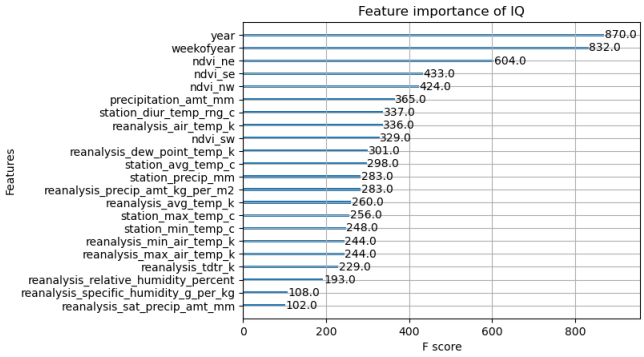

Fig. 9: Xgboost Feature Important of Iquitos

comprehensive approach led to a streamlined model, reducing the initial 24 features to a select few that significantly impact prediction accuracy. In this research, there are 18 features were selected using RFE in San Juan, while in Iquitos, 5 features were selected for model training.

In the LSTM model tailored for predicting outcomes in the city of Iquitos, implementing feature selection through Recursive Feature Elimination (RFE) significantly enhanced the model's accuracy, reducing the Mean Absolute Error (MAE) from 7 to 5.93. Conversely, for San Juan, the adoption of feature selection methods showed negligible effects on model performance.

In XGBoost,Use gradient boosting to analyzing these indicators for feature selection ,the features that contribute most significantly to the performance of the model can be identified,as illustrated in Fig. 8 and Fig. 9. Feature importance values represent contributions, which can improve the accuracy of the model and the performance efficiency of the model

4.3. Regression Model Results

In this research, it evaluated several regression models, including LR, RF, MLP, LSTM, XGBoost. Model performance was assessed using Mean Absolute Error (MAE)

4.3.1. Parameter Settings

This study focused on predicting dengue fever cases in two distinct cities: San Juan and Iquitos. By employing parameter tuning and grid search techniques, it tailored the LSTM and XGBoost models to each city's unique characteristics, optimizing for local prediction performance. The optimal parameter settings selected for each city are detailed in the Table 1. This customized approach ensures that the models deliver the highest accuracy possible in their respective urban contexts.

4.3.2. Model Comparsion

The efficacy of this dengue fever prediction models was evaluated using the Mean Absolute Error (MAE) metric, with the findings presented in the table below. For the city of San Juan, the XGBoost model outperformed other methods with the lowest MAE of 13.42, indicating a higher prediction accuracy. The Long Short-Term Memory (LSTM) network followed with an MAE of 17.4. The Multi-Layer Perceptron (MLP), Random Forest (RF), and Logistic Regression (LR) models yielded MAE values of 21.23, 26.24 and 27.4, respectively, suggesting a comparative variance in the prediction accuracy for this city.

In the case of Iquitos, the models demonstrated overall lower MAE values, suggesting better prediction performance

Model	Parameter	San Juan	Iquitos
LSTM	nb_epoch	50	100
	neurons	50	50
	batch_size	64	64
	lr	0.001	0.01
XGBoost	n_estimators	750	750
	lr	0.1	0.15
	max_depth	10	7
	subsample	0.75	0.75
	colsample_bytree	0.75	0.75

Table 1: Model Performance Comparison

than for San Juan. The LSTM model achieved the best performance with an MAE of 5.93, closely followed by the XGBoost model with an MAE of 6.00. The Logistic Regression model reported an MAE of 6.17, while the Random Forest model had an MAE of 6.84. The MLP model also showed a competitive MAE of 7.2 .

These results highlight the variability in model performance between the two cities and underscore the necessity of customizing model parameters to suit the epidemiological and environmental nuances of each locale. The detailed MAE values for each model are as Table 2.

City	Models	MAE
San Juan	XGBoost	13.42
	LSTM	17.4
	MLP	21.23
	RF	26.24
	LR	27.4
Iquitos	XGBoost	6.00
	LSTM	5.93
	MLP	7.20
	RF	6.84
	LR	6.17

Table 2: Model Performance Comparison

5. DISCUSSION

This discussion provides a critical assessment of this regression task's results, comparing and contrasting methodologies and outcomes with those of team members and prior research on this dataset. The analysis highlights the innovation and efficacy of specific approaches, pinpointing opportunities for enhancement and avenues for future investigation.

A) Comparison with Different Approaches

In San Juan, the XGBoost model demonstrated superior performance with the lowest MAE, suggesting its robustness in capturing complex patterns within the epidemiological data. In contrast, the LSTM, which excels in sequential data prediction, did not perform as expected, implying a possible overfitting scenario or a need for further hyperparameter tuning. The MLP, RF, and LR models, despite their differing approaches to prediction, yielded higher MAE values, indicating less accuracy in this context. It is worth considering that San Juan's climatic and environmental features may present a complex dynamic that XGBoost can capture more effectively than the other models.

On the other hand, Iquitos showed a different trend. The LSTM model emerged as the most accurate, closely mirrored by the XGBoost. This finding is particularly noteworthy as it underlines the LSTM's capacity to handle temporal sequences effectively, perhaps due to unique temporal patterns in the Iquitos data that align well with the LSTM's strengths. The performance of LR in Iquitos was also commendable, outperforming its own result in San Juan by a significant margin. XGBoost, while still among the top performers, did not lead as it did in San Juan.

The results emphasize the necessity for location-specific model tuning. The generalizability of models across different geographical locations is challenged by the variability in model accuracy as shown in this study. This underscores the importance of incorporating local data nuances in the modeling process, rather than relying on a specific approach. The findings have been interpreted with a recognition of the inherent complexities in disease prediction. The critical comparison suggests that while some models are versatile, there is no absolutely perfect solution. The results encourage a tailored approach to model selection and parameter optimization, considering the unique epidemiological environment of each target area.

This analysis not only assists in understanding the current predictive capabilities for dengue fever but also provides a foundational approach for similar studies in the field of infectious disease forecasting. It contributes to the ongoing conversation about the most appropriate and effective methods for disease outbreak prediction, a vital tool in public health planning and intervention.

B) Comparison with Previous Research

In this study, comparing the LSTM predictions of dengue fever incidence in Malaysia [6], with research conducted on the Iquitos city dataset. It applied various feature selection techniques to enhance the predictive accuracy of LSTM model. After refining the dataset through these methods, the LSTM model achieved a mean absolute error (MAE) of 5.93 in forecasting dengue fever cases in Iquitos. This comparison not only highlights the effectiveness of feature selection in improving model performance but also provides a benchmark against previous studies, offering insights into the geographical variability of model accuracy in dengue prediction.

Building on previous research that utilized XGBoost to predict the spread of dengue fever, this study improved upon

the existing framework by analyzing the data, employing feature selection techniques, and optimizing model parameters. These enhancements led to a modest increase in prediction accuracy for two countries, with a sequential improvement of 10-20% , the Mean Absolute Error (MAE) was reduced to 13.42 in San Juan and to 6.0 in Iquitos.

6. CONCLUSION

This study sets out to evaluate the capability of machine learning (ML) techniques in forecasting dengue fever instances, utilizing data from San Juan and Iquitos. An extensive application and comparative analysis of several regression models—including Linear Regression, Random Forest, Multiple Linear Regression, Long Short-Term Memory (LSTM) networks, and XGBoost—provided valuable insights into their predictive accuracies, underscored by the critical role of data preprocessing and feature selection.

The research indicates that the XGBoost model outshines others in San Juan, recording a Mean Absolute Error (MAE) of 13.42 and standing out as the premier model for projecting weekly dengue cases. In Iquitos, the LSTM model, when synergized with Recursive Feature Elimination for feature refinement, surpassed its counterparts with a MAE of 5.93, emphasizing its proficiency in capturing the distinctive seasonality of the data for reliable forecasts.

The study's significance stems from determining the most efficacious ML models for dengue prediction and elucidating the influence of data preprocessing and feature selection on model accuracy. This foundation encourages further research to refine these models, incorporate additional factors, and apply this discoveries to other dengue-prone regions.

Subsequent research should focus on enhancing dataset richness by including a broader spectrum of variables that may influence dengue incidence, such as socio-economic factors and genetic markers. Moreover, deploying these models in practical settings for predicting dengue outbreaks could be instrumental in public health, enabling prompt responses and efficient distribution of healthcare resources.

In summation, this work highlights the profound implications of machine learning in epidemiology, especially for dengue case prediction. The insights offered by this study contribute to the scholarly dialogue surrounding the use of ML in healthcare and provide actionable strategies for improving dengue surveillance and public health initiatives.

7. REFERENCES

[1] Maria G. Guzman and Scott B. Halstead, "Dengue: a continuing global threat," *Nature reviews. Microbiology*, vol. 8, no. 12 0, pp. S7–16, Dec. 2010.

[2] Naiyar Iqbal and Mohammad Islam, "Machine learning for dengue outbreak prediction: A performance evaluation of different prominent classifiers," *Informatica (Ljubljana)*, vol. 43, no. 3, pp. 363–371, 2019, Place: LJUBLJANA Publisher: Slovensko Drustvo Informatika.

[3] Gisella Luisa Elena Maquen-NiÃ±o, Jessie Bravo, Roger Alarcon, Ivan AdrianzÃ©n Olano, and Hugo Vega-Huerta, "Una revisiÃ³n sistemÃ¡tica de modelos de clasificaciÃ³n de dengue utilizando machine learning /a systematic review of dengue classification models using machine learning," , no. 50, pp. 5–, Publisher: AISTI Iberian Association for Information Systems and Technologies.

[4] Tzong-Shiann Ho and Ting-Chia Weng, "Comparing machine learning with case-control models to identify confirmed dengue cases," vol. 14, no. 11, pp. e0008843, Publisher: Public Library of Science.

[5] Dilip Kumar Choubey, Adweat Mishra, Sambeet Kumar Pradhan, and Naman Anand, "Soft computing techniques for dengue prediction," in *2021 10th IEEE International Conference on Communication Systems and Network Technologies (CSNT)*, June 2021, pp. 648–653.

[6] Abdulrazak Yahya Saleh and Lim Baiwei, "Dengue Prediction Using Deep Learning With Long Short-Term Memory," in *2021 1st International Conference on Emerging Smart Technologies and Applications (eSmarTA)*, Sana'a, Yemen, Aug. 2021, pp. 1–5, IEEE.

[7] Mr Ashish P. Joshi and Biraj V. Patel, "Data Preprocessing: The Techniques for Preparing Clean and Quality Data for Data Analytics Process," *Oriental Journal of Computer Science and Technology*, vol. 13, no. 2,3, pp. 78–81, Jan. 2021.

[8] Douglas C. Montgomery, Elizabeth A. Peck, and G. Geoffrey Vining, *Introduction to Linear Regression Analysis*, John Wiley & Sons, Feb. 2021, Google-Books-ID: tCIgEAAAQBAJ.

[9] Jersson X. Leon-Medina and Camacho-Olarte, "Monitoring of the refractory lining in a shielded electric arc furnace: An online multitarget regression trees approach," *Structural Control and Health Monitoring*, vol. 29, no. 3, Mar. 2022.

[10] Donald Salami, Carla Alexandra Sousa, Maria do Rosário Oliveira Martins, and César Capinha, "Predicting dengue importation into europe, using machine learning and model-agnostic methods," *Scientific Reports*, vol. 10, no. 1, pp. 9689, 2020.

[11] Ralf C Staudemeyer and Eric Rothstein Morris, "– Understanding LSTM – a tutorial into Long Short-Term Memory Recurrent Neural Networks," .