Machine learning-based dengue forecasting in San Juan and Iquitos using meteorological variables

Abstract—**Dengue is currently a disease of global concern, for which there is no specific drug or vaccine. In order to prevent the outbreak of dengue fever, this project developed a model for the number of dengue fever in San Juan and Iquitos. In this dataset, the information recorded in SAN Juan is from April 30, 1990 to April 22, 2008, and the information recorded in SAN Juan, Iquitos is from July 1, 2000 to June 25, 2010. Through many experiments, this project combines different machine learning models and data preprocessing methods. The Random Forest model in this project combined a two-week lag in the environmental data with an outbreak forecast to provide the best prediction for both cities. And the negative binomial regression model has better results in less time.**

I. Introduction

According to the WHO report[1], Dengue Fever is now endemic in more than 100 countries in Africa, the Eastern Mediterranean, Southeast Asia and the Americas. The largest number of dengue cases will be 6.5 million in 2023, with the tropics and subtropics being the worst affected. The region of the Americas has a total of 4.5 million cases and 2,300 deaths. Asian cases are also high, with more than 321,000 in Bangladesh and 111,400 in Malaysia.

There is no vaccine for dengue and treatment is limited. Therefore, it is necessary to prevent dengue fever in some areas and countries where dengue fever often occurs.

Drivendata provides dataset on the environment and the number of cases in SAN Juan and Iquitos.In this dataset, the information recorded in SAN Juan is from April 30, 1990 to April 22, 2008, and the information recorded in SAN Juan, Iquitos is from July 1, 2000 to June 25, 2010. Weekly environmental data were recorded during this period as well as the total number of cases.

The environmental data include the following 0.5x0.5 degree scale data: 1) Total precipitation, 2) Mean dew point temperature, 3) Mean air temperature, 4)Mean relative humidity, 5) Mean specific humidity, 6) Minimum air temperature, 7)Average air temperature, 8) Diurnal temperature range, 9) Pixel northeast of city centroid, 10) Pixel northwest of city centroid.Also con- tains , 11) Maximum temperature, 12) Minimum temperature, 13) Average temperature, 14)Total precipitation, 15) Diurnal temperature range, 16)Pixel southeast of city centroid, 17)Pixel southwest of city centroid, 18) Pixel northeast of city centroid, 19)Pixel northwest of city centroid 20) Total precipitation(0.25x0.25 degree scale).Some features are displayed on different scales but represent the same metric.

Through preliminary observation and understanding the significance of each column of data, there are 3 columns of time data, 4 columns of data related to vegetation indicators, 10 columns of data related to temperature indicators, and 6 columns of data related to humidity indicators in this data set.

Our research question is: Whether dengue total case per week can be accurately predicted based on time and different environment indicators. In this experiment, different machine learning models will be used to predict future dengue outbreaks based on the existing data.

This report begins with Section II related to previous research on Dengue fever using Machine Learning, then Section III follows to introduce the detailed process of preprocessing and modeling. After the project is established, Section IV is set to show the outcome graphs and charts from models. Besides, Section V will receive a further assessment by comparing each other in the Discussion section. In the end of the report, Section VI, including strong-weak analysis and outlook of future application will be given to models as a whole.

II. Literature Review

At present, machine learning models are widely used for epidemic prediction, of course Dengue Fever can also be predicted by combining machine learning models with environmental data. For example, Baquero [2] et al. used the data of rainfall, temperature and humidity in the city of Sao Paulo in Brazil, he combines different machine learning and statistical models to predict the dengue cases in this city, and compare the performance of each model. A research team which also comes from San Paulo, Brazil imports data of Brazil from 2009 to 2017 into a Random Forest model with feature selection when preprocessing[3]. Due to the complex environmental conditions in Brazil, they admit that their research has drawbacks in explainability, but it is undeniable that the Machine Learning method is a valuable addition to local epidemiologist's toolbox. Besides, two researchers, Fernanda Paula Rocha and Mateus Giesbrecht carried out research in São Luís do Maranhão, Brazil about Dengue fever and tried to make prediction on it by applying machine learning models like Multiple Logistic Regression, Decision Tree and Random Forest[4].

Given that large countries may bring multiple environmental conditions, research that is carried out in smaller regions is easier and of more accuracy.

In Singapore, some Chinese researchers tried many Machine Learning models including SVM and XG Boost in the data of Dengue fever patients and weather situation in Singapore in order to find the relationship between them by applying feature engineering. They drew conclusion that Dengue fever has an explicit seasonal pattern, and factors that have a strong relationship with mosquito's reproduction are more important in Dengue fever detection.[5]

Another Indian team also selected Singapore as the Dengue research target. They introduced Classification Algorithm to design the prediction model so that they could provide medical staff with reliable statistics in finding patients and strengthening the detection of rural areas.[6]

Two cities are selected in our Dengue prediction project: Iquitos in Peru and Saint Juan in Perto Rico, which have been analyzed by many researchers so far.

In Peru, acute Dengue fever is thoroughly monitored by a research team from Maryland.[7] Time and space condition is also taken into consideration in another research project.[8] In addition, an essay which states that the Dengue fever is related with Covid-19 is proposed by a local academic group, and it is also particularly convincing. [9]

In Perto Rico, there is a time and space analysis in Dengue fever conducted by Virology Department, United States Naval Medical Research. [10] Besides, Oxford University Press published a paper which compares Dengue fever with two common viruses, Zika Virus and Chikungunya to find its feature in Perto Rico. [11]

Our research is implemented under the guidance of these academic seniors. And various models will be imported to analyze Dengue fever in Iquitos and San Juan.

III. Methodology

We only divided the dataset into training and validation set since the test set has already been reserved by the competition organizer. The rankings and scores obtained by the submission in this competition will be used as the performance of the model on the test set.

General step: Split by city. In this task, the data comes from two different cities, Iquitos and San Juan. Although they are both in the tropics, they are located in the northern and southern hemispheres respectively. This may cause them to have different climates at the same time of the year and affect the outbreak period of dengue fever cases. Therefore, before other Exploratory Data Analysis, first splitting the training set and test set according to different cities may be a good choice. Also, two models are designed for sj and iq city separately to achieve better prediction results.

In this task, our goal is to find a relatively better preprocessing method and model that performs in this data set.Attentions are paid to different types of machine learning models and their corresponding preprocessing methods by team members separately. After all team members have their respective types of models optimized, further comparisons are carried out to draw the final conclusion.

A. Tree-based model

1) Dealing with missing value: In the dataset for Iquitos, the feature with the highest number of missing values has 37 missing entries. Given that this relatively small number and data in this task is time-related, linearly interpolating data becomes a relatively better option. But for another city, San Juan, the feature with the highest number of missing values "ndvi_ne" has 191 missing entries accounts over 20%. Therefore, dropping this feature and linearly interpolate other features is better for sj_data.

2) Fix Feature (Year): In the original data, it is found that the maximum value of the week_of_year feature is different each year. In some cases, data collectors will mark the first week of certain years as week 53 in the data set. In order to avoid possible problems caused by inconsistent week numbers, one was added to all the week numbers in years with 53 weeks, and then convert the original 53 weeks into the first week.

3) Feature Selection: The general idea of this part is to select correlated features to the label first, and then conduct correlation analysis between features. Correlation analysis between feature and label is performed first. The absolute value of the correlation between each feature with label is ranked to select and drop least related features. After selecting the features more related to label (0.1 is chosen as the threshold), the correlation analysis between features is visualized by heatmap. The multiple temperature data retained in the previous selection show strong correlations in this step. But directly selecting few of them may lose some information. As a result, a new feature "general_temp" was generated to replace these related temperature data.

4) Feature Engineering: As mentioned in the previous part, this part will generate a new feature general_temp based on the current temperature features to summarize their information. In order to avoid the influence of different scales when combining temperature features, the selected temperature-related features are first normalized to 0-1. Then new feature general_temp was created and also scaled to 0-1 by adding their scaled value then divided by the number of selected features.

5) Final Features and Scaling: Finally, after deleting features that have low correlation with the label and merging temperature features that are strongly correlated with each other, four environmental features were selected for training the model. "relative_humidity_percent", "specific_humidity_g_per_kg", "precip_amt_kg_per_m2" and "general_temp" are selected and scaling to 0-1 for modelling.

6) Modeling-Evaluation-Modeling loop: In fact, Deep Neural Network (DNN) was selected at the beginning. All time-related data were all dropped because time as time-related data cannot be directly used as input to the neural network. From the visual comparison of predicted and actual values, the DNN can not predict the dengue outbreak by time. This is not only the reason for switching to a tree-based model here, but also the reason why another team member training time series model in Part 3.3. Then the decision tree was further evaluated and random forest was applied, and in the end the random forest was also evaluated and feature-shifting was added. The specific operations and reasons in the modeling-evaluation loop will be explained in detail in part 4: Results.

B. Negative binomial regression SVR and XGBoost

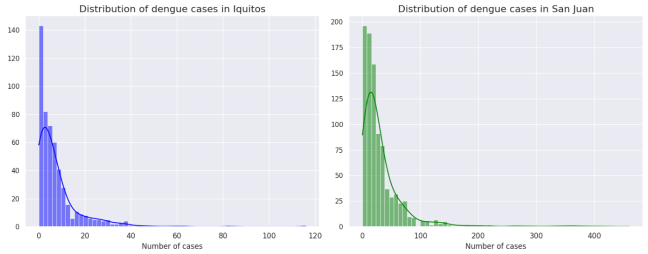

Fig. 1. data distribution of the total case of the two cities

1) Data preprocessing: In the beginning, the label and meteorological data are in different datasets, and they need to be merged. Since SAN Juan and Iquitos are in the north and south Hemisphere, they may not have the same temperature and rainfall at the same time, resulting in different dengue outbreaks, therefore the dataset needs to be divided by city. Then, it is essential to view the missing values in the dataset and see whether there are missing values in a number of columns. After observation, most of the columns don't have many missing values. The largest number of missing values is the ndvi ne of 191 missing. However, environmental data will not fluctuate significantly in a short period of time. Therefore, it is more appropriate to fill in the missing values instead of deleting them. Finally, the missing values are filled with the average of the previous value plus the next value. This is done to make the data smoother.

2) Feature Engineering: In the experiment, three different feature engineering methods are applied before training our models three times each. The first feature engineering is using the original data after being filled. The second feature engineering is to use an average sliding window of 2 units to modify the original data. The purpose of this is to make the data smoother and reduce the error of the data. The third feature engineering is to fill the data of the current week with the data of the previous two weeks. The reason for using this is that environmental factors can affect the growth cycle of mosquitoes with a time lag, which may subsequently influence dengue outbreaks. Moreover, before each training using SVR, it is necessary to scale the feature to avoid the impact of differing scales among features.

3) Feature Selection: In this experiment, a correlation analysis was conducted using a heat map to identify columns with redundant information. Features with correlation coefficient greater than 0.95 were removed. After that, another feature selection was guided by the correlation between the features and label. Features with a correlation coefficient less than 0.15 were considered to have low correlation with label and would be excluded from the training features. The second and third feature selection is the same logic, except that the second and third times modify the original data by rolling average and lagging 2 weeks of environmental data.

4) Modeling-Evaluation-Modeling loop: In this experiment, the negative binomial regression model is used first because the total case of this dataset belongs to the negative binomial distribution(see fig 1). Secondly, Support Vector Regression(SVR) is chosen, as it has a good performance in solving regression problems. Thirdly, the main reason for selecting eXtreme Gradient Boosting (eXtreme Gradient Boosting) as the third model is that it performs excellently in the robustness to outliers in the prediction accuracy. Overall, the three models were trained three times respectively. The first time, only feature selection was done without modifying the data, and in the second time, data is filled with the mean of it using a 2-unit sliding window, followed by a feature selection. For the third time, data of this week was filled with those of two weeks ago and then feature selection was conducted. Eventually, a set of 3-by-3 MAE results is generated, which is the final model to be analyzed.

C. Time Series Model

According to previous research carried out by other teams, Dengue fever has a strong seasonal feature, thus it is reasonable to apply time series model in this model.

1) Remove missing value: The first step of this model is dropping missing value because the filling of missing value is likely to affect the correlation calculation. After features are selected, rows with missing values will be recovered and processed at that moment.

2) Feature Selection: Feature Selection covers two stages. Firstly, the dataset needs to be split into 2 parts: Time-related features and numerical features because time features are not numerical and cannot be put into correlation heatmap. However, time features are of high correlation, therefore it is reasonable to reserve only one of them. As a result, "week-start-date" is introduced into the selected features, as it is an increasing feature with stable intervals.

For numeric features, a heatmap will be generated based on them, and features that have high correlation rate will be put into the same group. Features are separated into several groups. Only one feature in a group will be selected into the final combination. The number of groups is determined by a hyper-parameter: the correlation rate level, which decides whether two features are correlated.

In order to find the best hyper-parameter value, Linear Model which can reflect the performance of features is imported. The checking process is made on the split training set because the training set and testing set has similar data distribution, and before putting data into the model, data in the training set that generated from the former training set needs to be normalized. For every feature allocation, Linear Model provides the best combination of features that comes from each group and its score. By changing the value of hyper-parameter, combination with the best performance can be found and become the selected feature set that is put into the model.

3) Tear the data frame by city: The raw data comes from two different cities: Iquitos in Peru and San Juan in Perto Rico. These two cities are so far from each other that they are in different semi-spheres. Therefore, it is essential to cut the data set according to the city name.

4) Recover and deal with missing value: After the feature set is determined and split based on city name, the next step is filtering the raw data. Only columns with selected features are left. The new data frame has missing values because it comes from raw data, and it is necessary to handle these missing values before putting them into Machine Learning model. Missing values are replaced with median by using pandas.

5) Select Time Series Model: Now the data frame for machine learning has been preprocessed, it has some characteristics:

1. Time-related;
2. Huge capacity in rows;
3. Less columns compared with raw material;

Taking these characteristics into consideration, LSTM, a Machine Learning model related to time series, is a good choice because it performs well in long time data frame and can avoid vanishing gradient problem. Its main drawback is time-consuming, but the data frame has been reduced, so it has little effect on the final outcome. As a result, it is adopted as the time series Machine Learning model.

6) Make prediction and output as a csv: After the model received training, the testing set should be preprocessed in the same manner with training set. Afterwards, the LSTM model could generate predictions using the testing data set. Save the output into a new csv file.

IV. Results and Evaluation

In this section, each member will show the process how they optimized their respective models. Each team member will demonstrate the reasons and comparative effects of their respective method choices at each stage of this task. At the end of this section, each member will get their own best preprocessing method and model to prepare for combination and further comparison in section ”V. Discussion”.

Mean absolute error (MAE) is chosen as our evaluation matrix since we will follow the evaluation matrix specified by the competition organizer.

1) Deep Neural Network(DNN): After dividing the dataset into two cities, the commonly used DNN is first trained. Since time-related data cannot be directly used as input to the neural network, we decided to drop them directly. In the training process, K-flod (K = 5) was used for the validation set to verify the generalization performance of the model, and early stopping was used to avoid over-fitting. When the model does not improve on the validation set loss for more than 10 rounds, it will stop early. It can be seen from the comparison between the predicted value and the real value in the iq city in fig. 2.

Modeling-Evaluation loop:

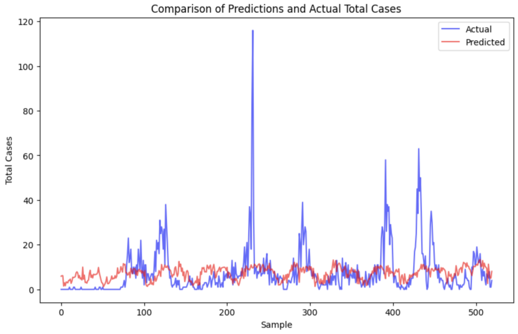

Fig. 2. Predicted and actual numbers of dengue cases of DNN in iq city

Performance: Validation set: MAE: 5.31 for iq city and MAE: 24.14 for sj city
Test set: MAE: 27.21 in the competition, Ranked #4193
Problem: From the visual picture of predicted and actual values, the DNN can not predict the dengue outbreak by time.
Analysis: All time-related features are dropped in order to ensure all inputs are of numeric type. However, further EDA proved that this regression task is related to time, especially the outbreak period of dengue fever that DNN cannot predict.
Possible solution: Train tree-based model or time series model to consider time related characteristics.

A. Tree-based model

1) Decision Tree: Compared to black box models, decision tree models are more trusted due to their interpretability and more types of inputs. [12] The deleted time features ’year’, ’weekofyear’, ’week_start_date’ are added back to the data. Fig. 3 shows the comparison between labels and predictions.

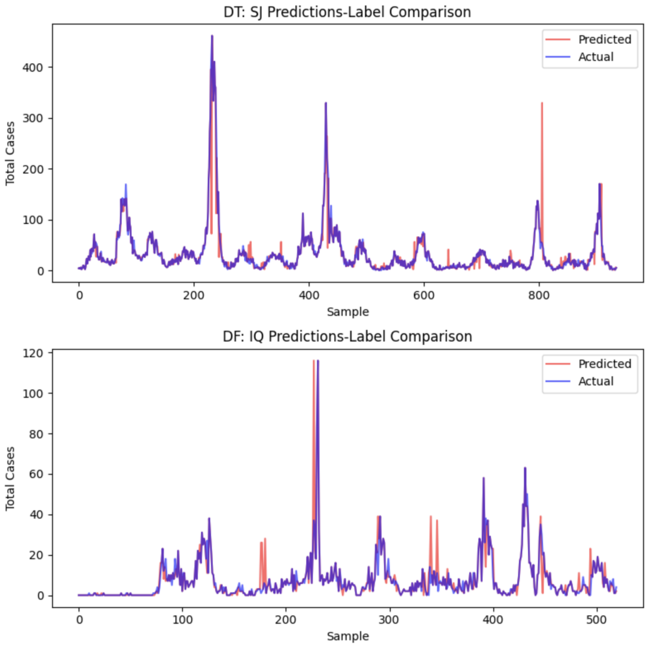

Fig. 3. Predicted and actual numbers of Decision tree in different city

Modeling-Evaluation loop:
Performance: Validation set: MAE: 6.06 for iq city and MAE: 15.49 for sj city
Test set: MAE: 30.45 in the competition, Ranked #5370
Problem: The decision tree model seems to be overfitting. Compared with the original DNN model, the performance on the verification set has significant improved, but the performance on the test set has declined.
Analysis: When faced with complex problems, decision tree will easily fall into overfitting because it will grow too deep and have too many nodes.[13] This is consistent with the results in visualization of prediction.
Also, incorporating years into training may also cause problems. For example, in iq city, the training set data consists of 2000 to 2010, while the test set years are from 2010 to 2013, which means that the year feature is unseen data for the trained decision tree model. This may make the tree model perform worse on the test set.
Possible solution: Deal with feature ”Year” and try random forest to avoid overfitting.

2) Further EDA and creation of outbreak-low period feature: As mentioned in the last section, adding year feature to training may be one of the reasons for overfitting. However, since the number of dengue fever cases is different in the same week of different years, it is unwise to delete the year data directly. After deleting the ”year” feature and conducting further EDA, we noticed that although the number of dengue cases per week was different in different years, the high and low periods of each city's annual incidence seemed to show a certain pattern. Therefore, we decided to introduce a new feature to capture this cyclical change.
For iq city: week 1 to 11 marked as sub-peak period, week 12 to week 20 marked as sub-trough period, week 21 to 36 marked as trough period and week 37 to 52 marked as peak period (Shown in Fig. 4).
For sj city: week 1 to 10 marked as sub-peak period, week 11 to week 22 marked as sub-trough period, week 23 to 30 marked as trough period and week 31 to 52 marked as peak period.

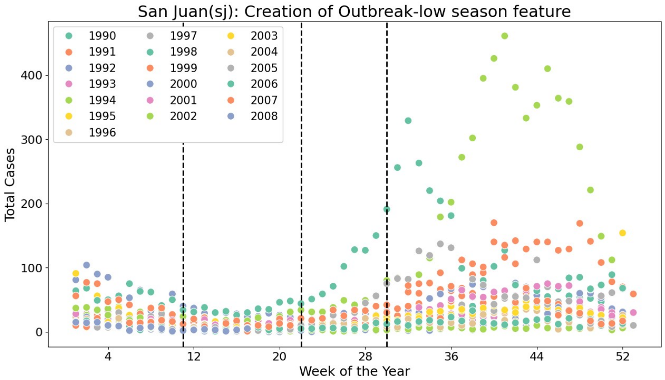

Fig. 4. Outbreak period division and creation new characteristics

3) Random Forest: Compared with the decision tree, random forest regression model which better adapt to the characteristics of this data set is trained. During the model training process, grid search method is applied to search for the optimal parameter combination.
Performance: Test set: MAE: 25.23 rank: The search process and comparison of prediction and label are presented in Fig.5:

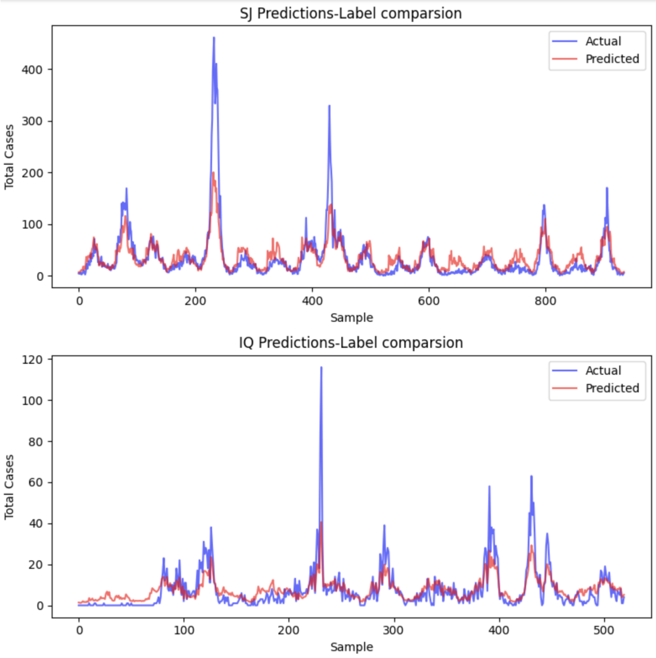

Fig. 5. Predicted and actual numbers of Random Forest in different cities

B. Negative binomial regression, SVR and XGBoost

1) Negative binomial regression: After dividing the data of two cities, it is found that the total case obeys the negative binomial distribution through the distribution analysis of the total case. Therefore it was decided to use a negative binomial regression model.
Modeling-Evaluation loop(3 times):
First time: Parameters : The optimal alpha parameter, alpha =1e-08, was obtained after several experiments by grid search.

Performance of validation set: MAE: 5.24 for iq city and 23.81 for sj city. General MAE: 26.12 on the test set.
Second time: Parameters : The optimal alpha parameter, alpha =1e-08, was obtained after several experiments by grid search.

Performance of validation set: MAE: 5.38 for iq city and 21.32 for sj city. General MAE: 26.43 on the test set.
Third time: Parameters : The optimal alpha parameter, alpha =1e-08, was obtained after several experiments by grid search.

Performance of validation set: MAE: 6.47 for iq city and 21.75 for sj city. General MAE: 25.90 on the test set.

2) SVR: **Performance(3 times): First time**: Parameters: The optimal parameters C= 0.01, epsilon=0.05, kernel=linear for sj and parameters C= 130, epsilon=0.03, kernel=linear for iq. Both city were obtained optimum parameter after several experiments by grid search.

Performance of validation set: MAE: 5.24 for iq city and 23.81 for sj city. General MAE: 28.39 on the test set.

Second time: Parameters: The optimal parameters C= 0.05, epsilon=0.3, kernel=linear for sj and parameters C= 130, epsilon=0.03, kernel=linear for iq. Both city were obtained optimum parameter after several experiments by grid search.

Performance of validation set: MAE: 4.63 for iq city and 20.24 for sj city. General MAE: 28.23 on the test set.

Third time: Parameters : The optimal parameters C= 0.05, epsilon=0.3, kernel=linear for sj and parameters C= 130, epsilon=0.03, kernel=linear for iq. Both city were obtained optimum parameter after several experiments by grid search.

Performance of validation set: MAE: 4.82 for iq city and 19.41 for sj city. General MAE: 27.61 on the test set.

3) *XGBoost:* **Performance(3 times):**

First time: Parameters : The optimal parameters learning rate = 0.01 max depth = 3,n_estimators = 80 for sj and learning rate = 0.01 max depth = 3, n estimators = 80 for iq. Both city were obtained optimum parameter after several experiments by grid search.

Performance of validation set: MAE: 5.24 for iq city and 25.57 for sj city. General MAE: 26.18 on the test set.

Second time: Parameters: The optimal parameters learning rate = 0.01 max depth = 1,n_estimators = 80 for sj and learning rate = 0.01 max depth = 1,n estimators = 100 for iq. Both city were obtained optimum parameter after several experiments by grid search.

Performance of validation set: MAE: 5.34 for iq city and 26.63 for sj city. General MAE: 26.46 on the test set.

Third time: Parameters: The optimal parameters learning rate = 0.01 max depth = 1, n_estimators = 80 for sj anlearning rate = 0.01 max depth = 1, n estimators = 100 for iq. Both city were obtained optimum parameter by grid search.

Performance of validation set: MAE: 5.48 for iq city and 25.75 for sj city. General MAE: 26.49 on the test set.

4) *Results of the three methods:*

TABLE I
MODELLING MAE SCORES

	Negative binomial regression	SVR	XGBoost
First time	26.12	28.39	26.18
Second time	26.43	28.23	26.46
Third time	25.90	27.61	26.49

Analysis: Through vertical comparison of the table I, it is found that the effect of negative binomial regression is always the best, the effect of XGboost model is the second, and the effect of svr is the worst among the three. The reason for the excellent performance of negative binomial distribution should be that the total case of this data set is subject to it. And SVR's bad performance can be attributed to the fact that the linear assumption of data may not conform to the distribution of actual data. SVR cannot capture complex nonlinear relationships, while XGboost can have better adaptation for datasets with negative binomial distribution.

Through the horizontal comparison, it can easily be found that lagging the environmental index data of the first two weeks will greatly improve the results, which is speculated to be due to the lag of environmental factors that affects the breeding number of mosquitoes and thus the number of dengue cases.

Problem: By observing Figure 6, negative binomial distribution can have a good performance in the non-outbreak period, but there is underfitting linearity in the outbreak period of dengue fever. As a result, it is necessary to separate the burst period from the non-burst period for the convenience of training in future work.

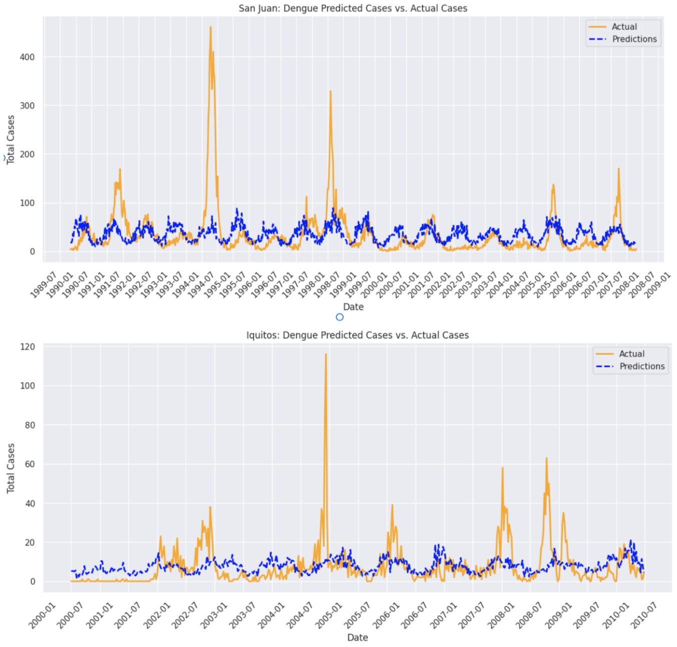

Fig. 6. The best effect predicted value and the true value of 9 experiments (negative binomial regression)

5) *Random Forest:*

C. *Time Series Model(LSTM)*

1) *Feature Selection Method:*

The hyper-parameter in feature selection was set as the level of belonging that can be attributed to correlated groups, which equals to the number of features remained in the selected feature set. For every possible quantity of features in a set, a best combination needs to be figured out.

For instance, when the hyper-parameter is 0.75, which means features that have a correlation ratio over 0.75 with each other will be put into the same group. By taking this, 19 features are allocated into 10 groups, and only one feature will be picked into the selected feature set in each group. This situation generates 192 possible combinations of feature set, so it is important to catch the combination with the best performance.

The tool used in this process is Linear Regression, and its score decides which is the best combination for a certain hyper-parameter (or number of features in selected feature set).

2) Hyper-parameter Selection: Following the last section, the best combination of selected feature set for every hyper-parameter is claimed and the next step is making comparison among them. The trend of performance with the value of hyper-parameter is increase first and then decrease, reaching a maximum at around 0.75 with 10 elements in the selected feature set.

3) LSTM-related parameter selection: This part focuses on the improvement of LSTM model's performance. For LSTM, parameter that can influence the result most dramatically is the time step. If time step is set unsuitably, the model can easily become overfitting or underfitting. Taking Iquitos as an example, the model become overfitting with timesteps = 3 and underfitting when timesteps = 15.

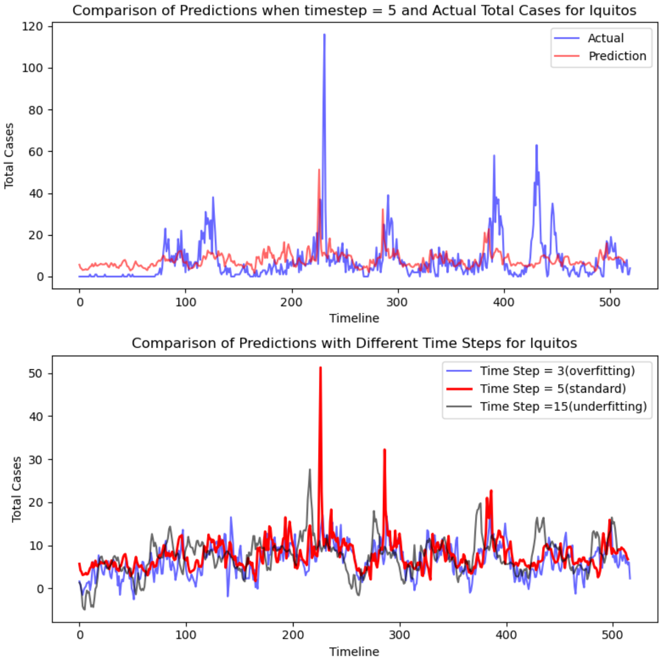

Fig. 7. Time Step optimization in Iquitos

Besides, the best time step differs according to the city location. For Iquitos, the best time step value is 5, while it is 15 in San Juan. By introducing LSTM properly, the seasonal rule of Dengue fever is magnified to the level where it can work as it did in realty.

V. Discussion

This study demonstrated the potential of machine learning models to predict dengue cases in San Juan and Iquitos. Before comparing it with other models from the literature, the final model will be established by conducting a comparative analysis within the team and integrating the optimal models discussed in part 4. The evaluation will be based on the MAE of the test set and the training time.

TABLE II
Final modelling MAE Scores

	Negative binomial regression (with lagged)	**RF (seasonal feature)**	**LSTM**
MAE	25.90	25.23	26.18
Training time (s)	0.41	3714.20	74.3

For the **pre-processing** part: During the group discussion, it was found that in addition to basic linearly interpolate missing data, feature selection and feature engineering, two unique preprocessing methods were worth paying attention to. Dengue outbreaks tend to occur a few weeks after environmental changes, since the impact of environmental factors on dengue fever exhibits a lag effect. It is necessary to add time-lagged features. Creating a new seasonal feature is also necessary to avoid the model overfitting specific year and weeks.
For the **modeling** part:

After taking the training time and prediction accuracy into consideration, negative binomial regression, which offers the highest training speed, and the random forest with relatively higher accuracy. The final results are obtained by combining different methods of feature engineering and feature selection.

After conducting many iterations and training the model, it was found that the two unique preprocessing methods mentioned in the preprocessing section, creating new outbreak period features and adding time lagged features, are both successful attempts and benefit for model training.

For the modeling part, negative binomial regression has the characteristics of simple training and low model complexity. However, just due to its overly simplistic structure, further optimization of the model proves challenging, especially when the distribution of the test set is difficult to ascertain.

Although the random forest model was selected as the final model by our group based on its performance on the test set, its long training time and possible overfitting problems are still a big challenge.

Finally, after combining all preprocessing methods and retraining the random forest model, we obtained a MAE score of 24 in the competition, which ranked 849 out of 14028 participants.

The project made by Fernanda Paula Rocha and Mateus Giesbrecht has a similar topic with this model and it is designed for São Luís do Maranhão[4], a city with similar latitude to Iquitos, therefore it is reasonable to introduce this project and make comparison with our model. The São Luís project features a complete preprocessing method and unique technology that can reduce the effect of overfitting. However, they did not make the most of the time feature, which is the key advantage of our model, which has an efficient and realistic preprocessing method. Comparing with that project, the model takes seasonal and time lagging feature into consideration, thus it achieves a better performance in making predictions.

VI. Conclusions and future work

In conclusion, after comprehensively comparing multiple preprocessing methods and models, the results show that the random forest model has the ability to effectively predict the future outbreak of dengue fever in SAN Juan and Iquitos cities. In data preprocessing stage, the creation of new features 'outbreak seasonal', along with the lagging of environmental features, has also proven to significantly enhance model performance.

However, our current application of these two preprocessing methods still needs closer integration. Future researches should consider the potential effects of different periods on the length lag times, which is also more consistent with the differences in the hatching times of insects as dengue fever media in different seasons. For example, during the summer when insect incubation periods are shorter, the length of lag times for environmental features should also be shorter to better reflect these biological changes.

On the other hand, although meteorological data have been completely collected and incorporated into model training, it is still insufficient to accurately predict the outbreak of dengue fever. Humanities and social information should also be an important part of the dengue prediction model. The size of the urban population, economic development, urbanization rates, and medical and health conditions will also have a significant impact on the outbreak of dengue fever. Collect more data in humanities and social area and adding them into our model in the future will enhance the robustness and predict dengue fever outbreaks more accurately. This will help alert local medical departments to prepare in advance to deal with dengue fever and save more lives.

References

[1] World Health Organization. Dengue and severe dengue. http://www.who.int/mediacentre/factsheets/fs117/en/, 2024.

[2] Oswaldo Santos Baquero, Lidia Maria Reis Santana, and Francisco Chiaravalloti-Neto. Dengue forecasting in são paulo city with generalized additive models, artificial neural networks and seasonal autoregressive integrated moving average models. *PloS one*, 13(4):e0195065, 2018.

[3] Kirstin Roster, Colm Connaughton, and Francisco A Rodrigues. Machine-learning–based forecasting of dengue fever in brazilian cities using epidemiologic and meteorological variables. *American Journal of Epidemiology*, 191(10):1803–1812, 2022.

[4] Fernanda Paula Rocha and Mateus Giesbrecht. Machine learning algorithms for dengue risk assessment: a case study for são luís do maranhão. *Computational and Applied Mathematics*, 41(8):393, 2022.

[5] Na Tian, Jin-Xin Zheng, Lan-Hua Li, Jing-Bo Xue, Shang Xia, Shan Lv, and Xiao-Nong Zhou. Precision prediction for dengue fever in singapore: A machine learning approach incorporating meteorological data. *Tropical Medicine and Infectious Disease*, 9(4):72, 2024.

[6] Rajeev Kapoor, Sachin Ahuja, and Virender Kadyan. Machine learning based classification algorithm for classification of dengue (dengue fever-df, dengue harmonic fever-dhf, serve dengue-sd). *ECS Transactions*, 107(1):4659, 2022.

[7] Eric S Halsey, Maya Williams, V Alberto Laguna-Torres, Stalin Vilcarromero, Victor Ocana, Tadeusz J Kochel, and Morgan A Marks. Occurrence and correlates of symptom persistence following acute dengue fever in peru. *The American journal of tropical medicine and hygiene*, 90(3):449, 2014.

[8] G Chowell, CA Torre, C Munayco-Escate, L Suarez-Ognio, R Lopez-Cruz, JM Hyman, and Carlos Castillo-Chavez. Spatial and temporal dynamics of dengue fever in peru: 1994–2006. *Epidemiology & Infection*, 136(12):1667–1677, 2008.

[9] Rubí Plasencia-Dueñas, Virgilio E Failoc-Rojas, and Alfonso J Rodriguez-Morales. Impact of the covid-19 pandemic on the incidence of dengue fever in peru. *Journal of Medical Virology*, 94(1):393–398, 2022.

[10] Gavino Puggioni, Jannelle Couret, Emily Serman, Ali S Akanda, and Howard S Ginsberg. Spatiotemporal modeling of dengue fever risk in puerto rico. *Spatial and Spatio-temporal Epidemiology*, 35:100375, 2020.

[11] Burke A Cunha, Anna Apostolopoulou, Thulashie Sivarajah, Natalie C Klein, et al. Facial puffiness in a returning traveler from puerto rico: chikungunya, dengue fever, or zika virus? *Clinical Infectious Diseases*, 63(9):1264–1265, 2016.

[12] Leilani H. Gilpin, David Bau, Ben Z. Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. Explaining explanations: An overview of interpretability of machine learning. In *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 80–89, 2018.

[13] Rafael García Leiva, Antonio Fernández Anta, Vincenzo Mancuso, and Paolo Casari. A novel hyperparameter-free approach to decision tree construction that avoids overfitting by design. *IEEE Access*, 7:99978–99987, 2019.