Erdős Institute Deep Learning May-Summer 2024 Executive Summary

Climate-Based Forecasting of Dengue Dynamics

Abdullah Al Helal, Haridas Kumar Das, Chun-hao Chen, Feng Zhu

Overview

Dengue outbreaks have become a global concern, affecting many regions such as the Americas, Africa, the Middle East, Asia, and the Pacific Islands. Over the past two decades, there has been a notable rise in dengue cases worldwide, with significant impacts observed in countries like Brazil and Bangladesh. Moreover, in the United States, local dengue transmission has been reported in a few states, including Florida, Hawaii, Texas, Arizona, and California. Numerous studies have demonstrated the correlation between climate factors—such as temperature and rainfall—and dengue, Zika, chikungunya, and yellow fever transmission. Specifically, elevated temperatures have been linked to an increased dengue infection risk, while extreme rainfall events have been shown to decrease this risk. In this project, we deploy deep learning time series analysis to analyze climate, epidemiological, and social data in order to forecast dengue dynamics.

Objective

In this project, we develop deep learning algorithms to analyze climate, epidemiological, and social data in order to forecast dengue dynamics, focusing on the analysis of Bangladesh.

Evaluation Methodology

Data Preprocessing

The Open-Mateo climate data, the Bangladesh dengue epidemiological data, the Google Trends social data were preprocessed using the following steps:

- The daily climate data was converted into monthly data by taking the mean.
- The number of cases from the daily epidemiological data was accumulated into monthly data.
- The monthly social data for 20 dengue related queries were converted into one monthly feature by taking the mean.

Exploratory Data Analysis

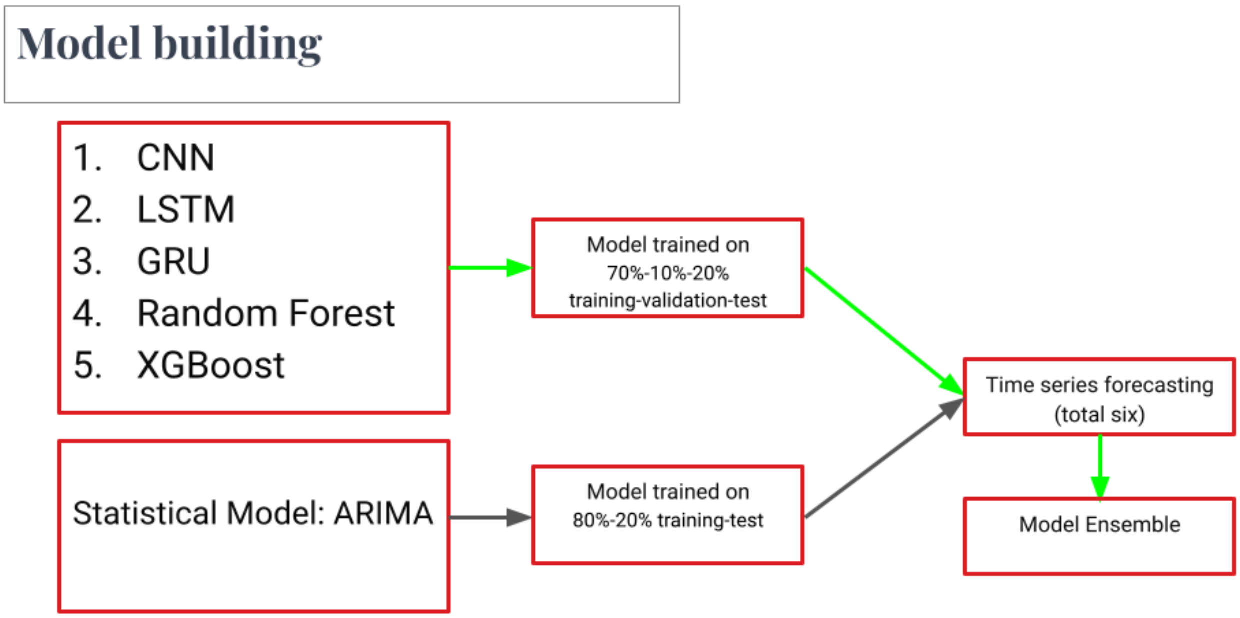
- There are 5 features, all numeric. There are 192 observations during the years 2008 to 2023.

Evaluation

- RMSE between the predicted and actual number of dengue cases was used as the evaluation metric.

Model building

- Five deep learning models were implemented to forecast dengue dynamics: CNN, LSTM, GRU, Random Forest, and XGBoost. One statistical model was also implemented: ARIMA.
- All deep learning models were used to forecast for the year 2024 using the same split of 70%, 10%, and 20% training-validation-test samples. The statistical model used a split of 80% and 20% training-test samples.



Ensemble

- Consider that we have RMSE scores R_i from M different models. We define the penalty for each model as:

$$\hat{w}_i = \frac{1}{R_i}.$$

- We normalize these penalties to obtain the final weights w_i, which are calculated as

$$w_i = \frac{\hat{w}_i}{\sum_{i=1}^{M} \hat{w}_i}.$$

- To compute the final ensemble prediction: Let y_i represent the prediction vector from the i-th model. The final ensemble prediction vector $y_{ensemble}$ is calculated as:

$$y_{\text{ensemble}} = \sum_{i=1}^{M} w_i \mathbf{y}_i.$$

Results

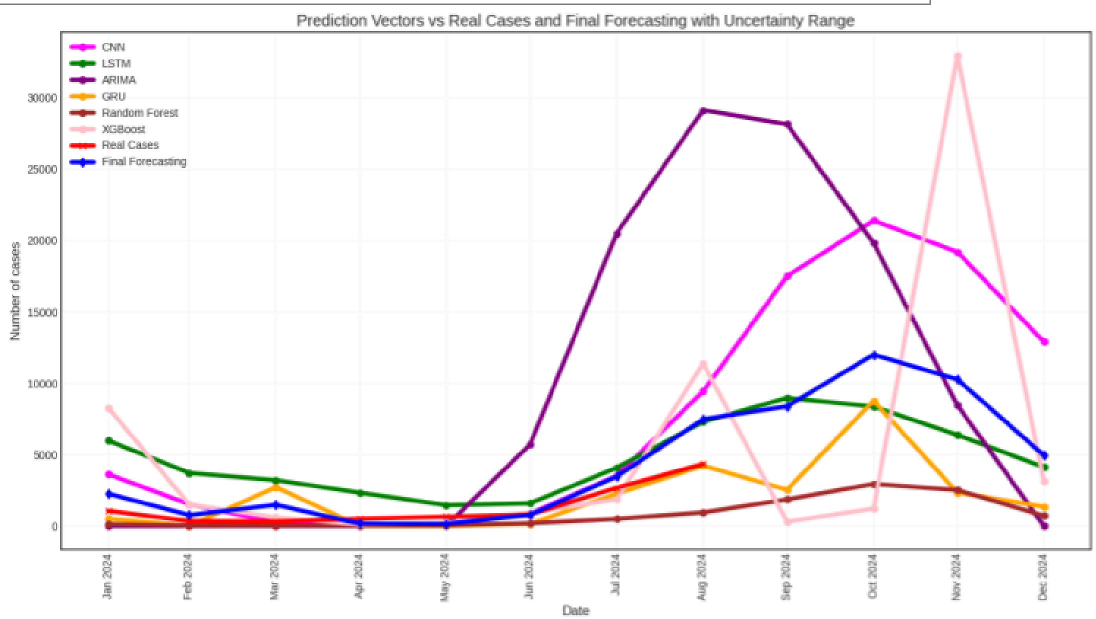
Evaluation on test set

- GRU resulted in the smallest test RMSE at 1943.58, CNN being a close second smallest at 2880. ARIMA resulted in the largest test RMSE at 17116.61. For reference, the highest monthly case was 79598 in September 2023.

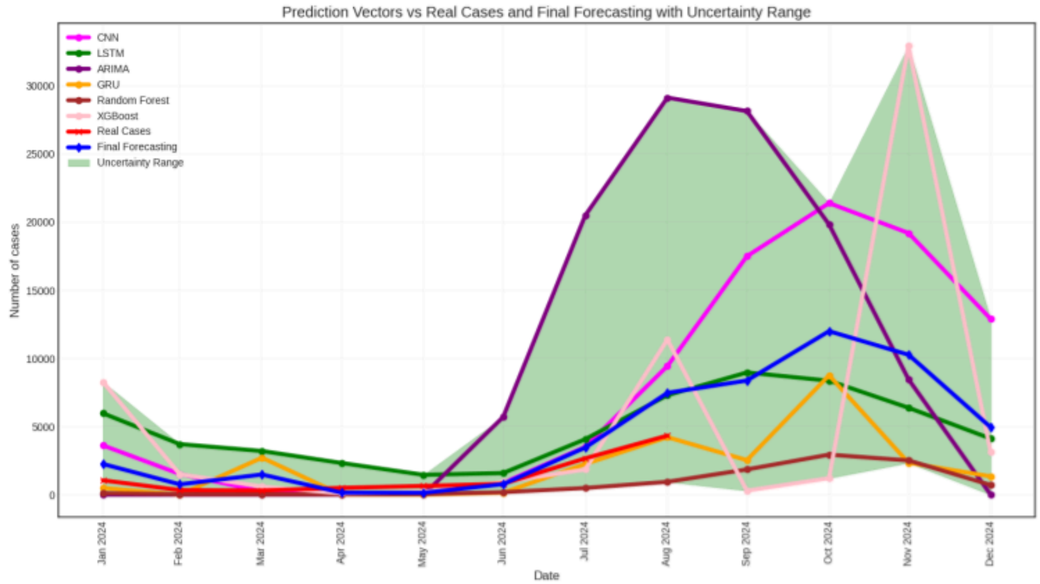
Forecasting

- We used each of the individual models, as well as a combined ensemble model (with weights indicated below), to produce a forecast for 2024. This involved forecasting case numbers for 12 data points beyond the dataset we have, as well as forecasting the climate and Google Trend covariates for these 12 months as part of this case number forecasting process using the following ensemble.
- It seems difficult to correctly forecast the magnitude of the peak, but more reasonable to expect models to correctly identify when the peak will be, and more generally identify trends in the case numbers.

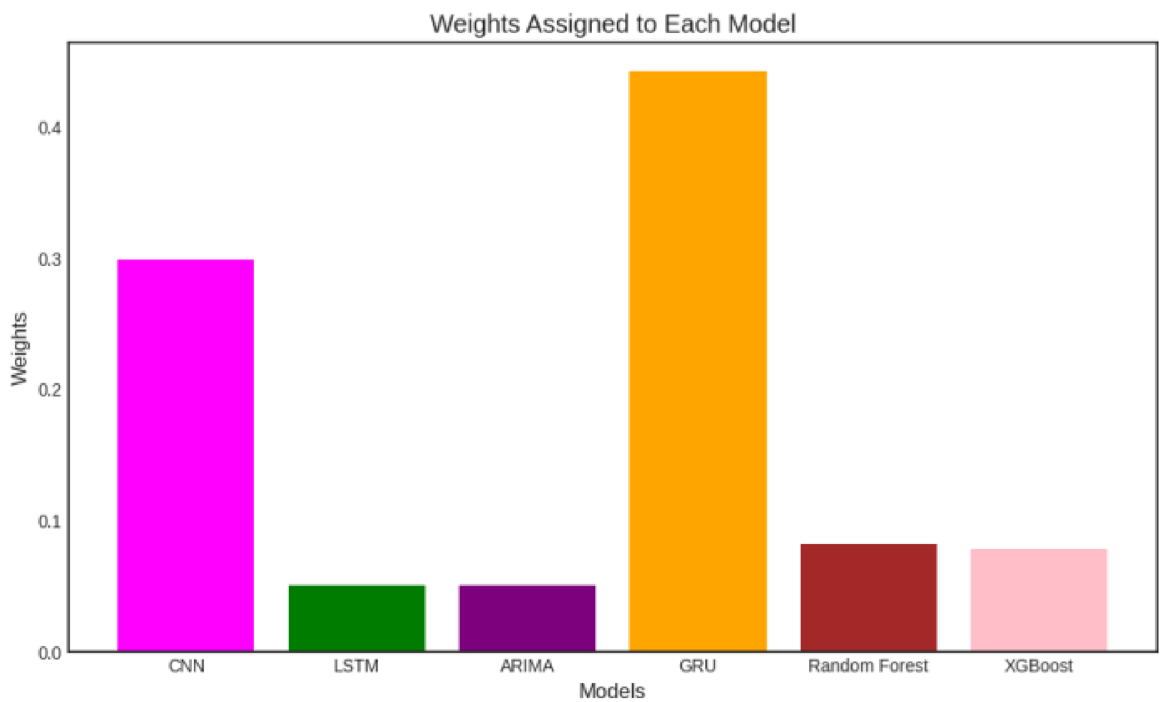
Ensemble Model : Forecasting



Ensemble Model : Forecasting



Model performance



Limitations

- One area for improvement in modeling lies in the potential for the model to overlook complex relationships between climate, epidemiological, and social data, which can result in less accurate predictions.

Conclusion and Future Directions

Conclusion

- Six deep learning and statistical models were deployed to forecast dengue dynamics in Bangladesh. Two methods—GRU and CNN—reasonably were able to forecast dengue cases for the year 2024 in terms of both test error and visual trend of forecasting in the unknown future.

Future Directions

- It remains to be explored if separating seasonal and trend components in the data improves model performance.
- Forecasting can be improved by using a bigger dataset.
- Future work will involve implementing hybrid deep learning models and sophisticated probabilistic time series forecasting algorithms.