DIARRHOEA OUTBREAK PREDICTION USING CLIMATE CHANGE

Introduction

Diarrhoea is a leading cause of morbidity and mortality on a global scale. In South Africa, it is the 3rd leading cause of death for children under five and 8th most preeminent cause of death in individuals of all ages. The region is a climate hotspot and will experience increase in frequency and magnitude of extreme events such as drought, heat waves etc. These events have the potential to increase communicable diseases such as diarrhea. Currently, possible threat is usually identified when there is an increased rate of hospitalizations due to the disease. Thus, outbreaks are predicted only when there is an epidemic. Using Machine Learning (ML), we may be able to establish an early-warning system for diarrhoea outbreak in the various provinces incorporating climate information.

Aim

The aim of this study is to use selected climate variables to make forecast about possible number of diarrhoea cases in the 9 South African Province. South Africa comprises of several climate variables. However, for this study, we considered the most widely used in climate impact studies which are Temperature, Precipitation Humidity, Pressure, Evaporation and Wind.

Objectives

The main objective conceived for this study is to detect which supervised machine learning techniques such as Convolutional Neural Networks (CNN), Long-Short Memory Networks (LSTM) and Support Vector Machines (SVM) performs best in terms of accuracy when predicting possible number of diarrhoea cases given a range of datasets.

Data & Methodology

The datasets used include daily sales records (2008-2018) of loperamide (anti-diarrheal medication) and daily record of climate factors of interest across the 9 provinces.

To determine the best performing ML algorithm for predicting possible number of daily diarrhoea cases (i.e. prediction for one day lead time), we compared the Root Mean Square Error (RMSE) from the predictions made by the three ML algorithms in different scenarios. The scenarios were broken down into different experiments which are:

- **Experiment I**: predictions with the original data and Grid search parameter tuning.
- **Experiment II**: predictions with augmented data (i.e. augmenting the original with synthetic data where synthetic data was generated with Generative Adversarial Networks (GANs)) and Grid search parameter tuning.
- **Experiment III:** predictions with augmented data used in Experiment II and an Evolutionary strategy called Relevance Estimation and Value Calibration (REVAC) to tune the parameters of all ML models.

In all the experiments listed above, the algorithm with the lowest RMSE errors indicate better prediction accuracy.

Results

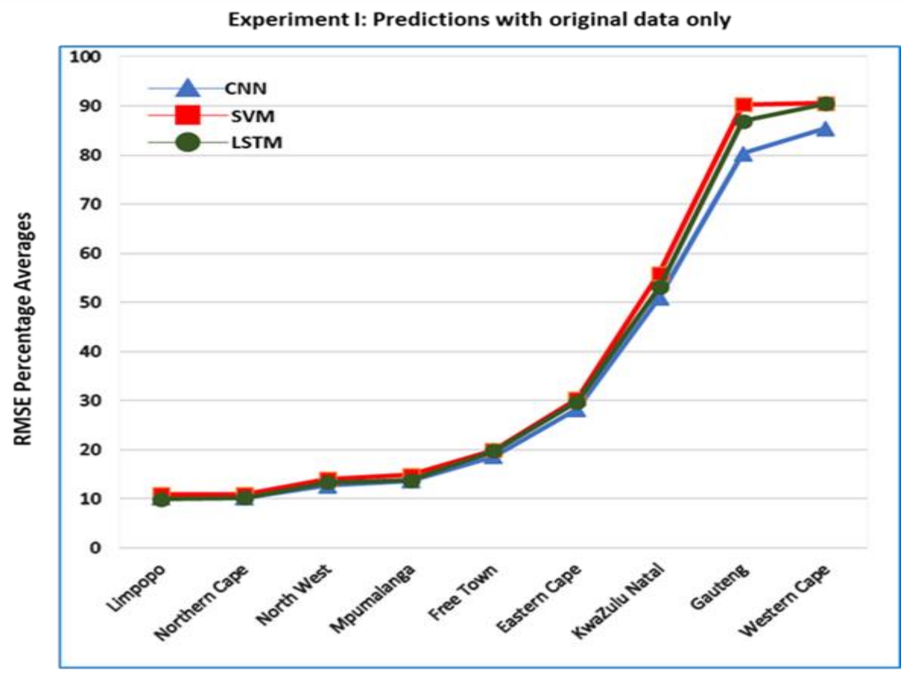

Figure 1: CNN, SVM, LSTM average RMSE when making predictions for possible number of daily diarrhoea cases in Experiment I. Low RMSE averages indicate better performance.

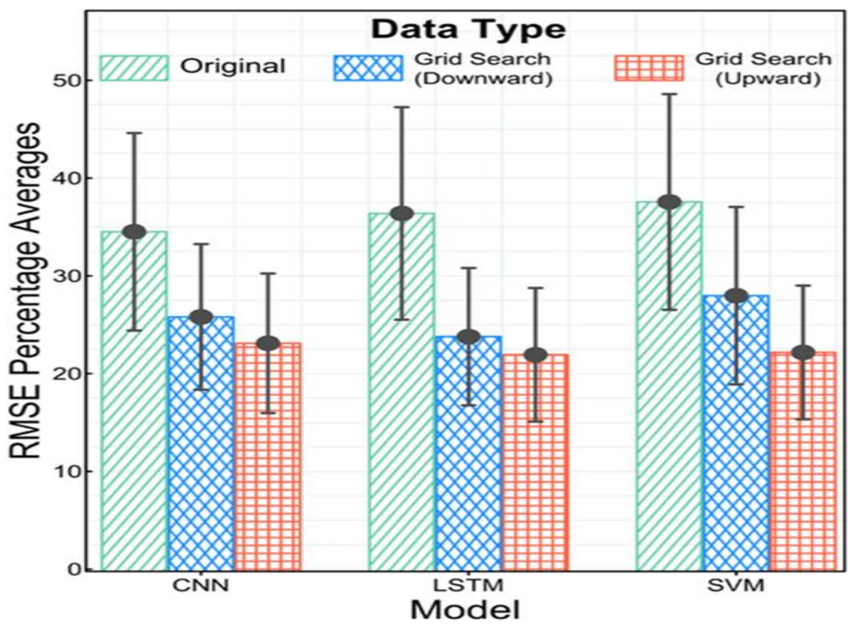

Figure 2: A comparison of CNN, LSTM, SVM average RMSE over all provinces with the original data in Experiment I and the augmented data in Experiment II. In Experiment II, data was augmented in two directions (upwards and downwards). Grid search tuning was used in both experiments. . Low RMSE averages indicate better performance. The arrows represent the corresponding widths of twice the standard error.

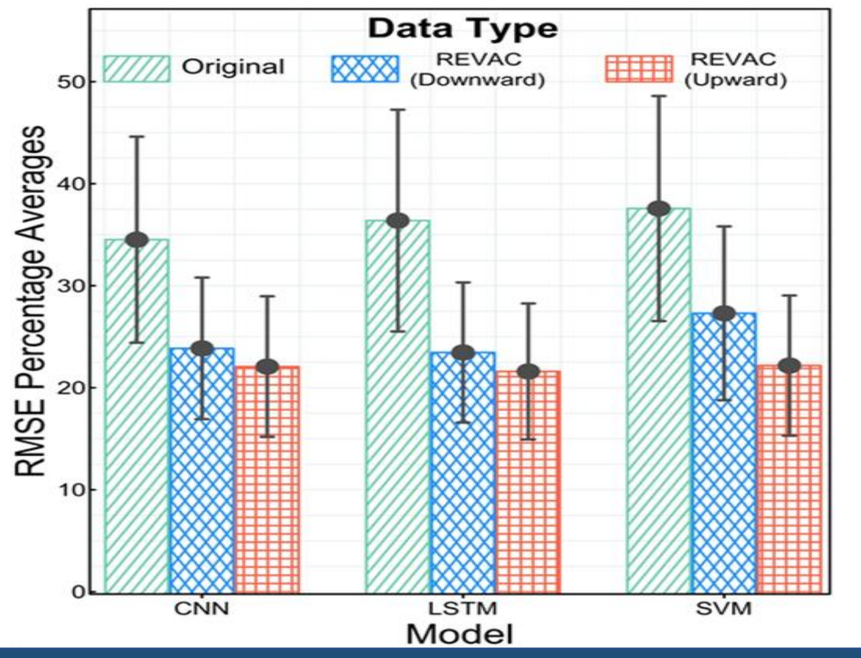

Figure 3: A comparison of CNN, LSTM, SVM average RMSE over all provinces with the original data in Experiment I and the augmented data in Experiment III. In Experiment III, data was also augmented in two directions (upwards and downwards). REVAC search tuning was used in Experiment III. Low RMSE averages indicate better performance. The arrows represent the corresponding widths of twice the standard error

Conclusion

The results from the three experiments showed that the three ML methods (CNN, LSTM and SVM) were appropriate for predicting daily diarrhoea cases with respect to the selected climate variables in each South African province. They were all able to yield low and similar RMSE. However, the level of accuracy for each model varied across the different experiments. In Experiment I, CNN outperformed, In Experiment II & III, LSTM outperformed. Overall, the deep learning models performed best in all Experiments. In addition, the use of data augmentation improved the performance of all ML models.

University of Cape Town
Private Bag X3
Rondebosch 7701
South Africa

Tassallah Amina Abdullahi
Supervised by: Dr. Geoff Nitschke