Predicting Diarrhoea Outbreak with Climate Change

A dissertation submitted in satisfaction of the requirements for the degree

Master of Science in Computer Science

University of Cape Town

by

Tassallah Amina Abdullahi

Supervised by:

Dr. Geoff Nitschke

December 2020

I know the meaning of plagiarism and declare that all of the work in this document, save for that which is properly acknowledged, is my own.

Acknowledgments

First and foremost, I thank Allah (the most gracious and most merciful), for giving me the strength and ability to complete this master's dissertation.

I owe sincere gratitude to my supervisor Dr. Geoff Nitschke, for his kindness, support and guidance provided to me throughout the course of my program. His belief in me has pushed me beyond boundaries that I did not know existed in me. I will do my best to put what I have learnt from you into good use.

I would like to say thank my colleagues and friends in the Computer Science Department. Thank you for always being there to chat and lend a helping hand. I am also grateful for the financial support that I received from the Mandela Institute for Development Studies and the NRF. Special thanks to Dr. Neville Swejid for his support and assistance in facilitating the provision of data for this study. I am also grateful to my lecturers at UCT for their tutelage.

Finally, I would like to express my love and acknowledge the support of my husband, Dr. Shakirudeen Lawal, my rock; and my son, Hashim Lawal. I say thank you for all your words of encouragement, your love, patience, and perseverance. I also want to thank my parents, siblings, and other family members for their support and prayers. I cannot thank my mom, Safiya Abdullahi enough for all the time she took to babysit during my program. Without the support of you all I would not have been able to complete this study.

Abstract

Climate change is expected to exacerbate diarrhoea outbreak in South Africa, a leading cause of morbidity and mortality in the region. In this study, we modelled the impacts of climate change on diarrhoea with machine learning methods. We applied two deep learning techniques, convolutional neural networks (CNNs) and long-short term memory networks (LSTMs); and a support vector machine to predict daily diarrhoea cases over the different South African provinces by incorporating climate information. Generative Adversarial Networks (GANs) was used to generate synthetic data which was used to augment the available dataset. Furthermore, relevance estimation and value calibration (REVAC) was used to tune the parameters of the machine learning algorithms to optimize the accuracy of their predictions. Sensitivity analysis was also performed to investigate the contribution of the different climate factors to the diarrhoea prediction model.

The results of the study showed that all three ML methods were appropriate for predicting daily diarrhoea cases with respect to the selected climate variables in each South African province. The ML methods were all able to yield low and similar RMSE. However, the level of accuracy for each model varied across different experiments, with the deep learning models outperforming the SVM model. Among the deep learning techniques, the CNN model performed best when only real-world dataset was used, while the LSTM model outperformed the other models when the real dataset was augmented with synthetic data. Across the provinces, the accuracy of all three ML algorithms improved by at least 30% when data augmentation was implemented. In addition, REVAC improved the accuracy of the CNN model by more than 12% in KwaZulu Natal province. However, the percentage increase in accuracy of the LSTM model was less than 4% in Western Cape province when REVAC was used. Our sensitivity analysis revealed that the most influential climate variables to be considered when predicting outbreak of diarrhoea in South Africa are precipitation, humidity, evaporation and temperature conditions. The result of this study is important for the development of an early warning system for diarrhoea outbreak over South Africa.

Contents

List of Figures

List of Tables

Chapter 1

1. Introduction

Diarrhoea is a clinical syndrome which changes the normal bowel movement by increasing the watery content and frequency of stools [1]. Major causes include unhygienic eating habits, gastrointestinal infections and diseases caused by bacterial, parasitic, and viral organisms [1], [2]. In most cases, diarrhoea is deadly for children under the age of five, yet adult mortality from diarrhoea is not unusual especially when there is a widespread occurrence of diarrhoeal related illnesses [2]. It is a major health concern and has remained the second leading cause of global morbidity and mortality [1], [3], [4]. Each year, over 2.5 million deaths attributed to diarrhoea are recorded worldwide [4]. Estimates suggest that most cases are concentrated in Sub-Saharan Africa and South Asia as they account for more than 80 percent of total world records [3], [4]. The increase in the number of diarrhoea cases during certain periods indicates that diarrheal diseases vary greatly with seasons, and global climate change is expected to increase its risk [5].

Extreme weather events such as droughts and heatwaves due to climate change affect human health directly or indirectly and as a result, the continued impacts will be one of the challenges in controlling infectious diseases in the future, especially in developing countries [5]. Investigations have shown these events tend to cripple public and environmental health thus several major killer diseases including diarrhoea related ones are projected be on the rise [5], [6]. In addition, climate factors such as temperature, rainfall may also contribute to changes in the incidence and severity of diarrhoea [6]. Nevertheless, diarrhoea is both preventable and curable [1], [2] .

However, the treatment and prevention of diarrhoea with vaccines, antibiotics, and anti-diarrhoeal medications (such as Loperamide) is a burden on public health system particularly in developing countries [2]. For example, the use of rotavirus, cholera, and typhoid vaccines to prevent diarrhoea is costly to the government; and the cure can be hardly afforded by low income families [7]. Therefore, it is recommended to develop and strengthen public health systems that aid in reducing the incidence and severity of diarrhoeal diseases [5]. One way to achieve this is develop a model that uses climate records to predict diarrhoea outbreak in advance. This information can be used for public health surveillance as it will offer timely detection and prompt notification for the control of diarrhoea outbreak. It will also enable government organizations and healthcare providers to take necessary action and put together intervention strategies to mitigate risks related to diarrhoea thus, minimizing the costs of delivering medical care related to

it. Machine Learning (ML) algorithms can help in detecting anomalies by reviewing the volume of data collected in health centres.

In recent years, various ML techniques such as artificial neural networks (ANNs), support vector machines (SVMs) and random forests (RF) have been used in developing predictive and diagnostic models for complex problems [8]. They have also been widely used in the medical field for diagnosis and disease prediction [8], [9] . For instance, in India, an SVM was used to predict malaria disease outbreak, the system was able to give about 15-20 days lead time for early intervention [10]. ANNs were also used for predicting the cancer outcome of an individual [11]. Apart from being applicable to complex problems, studies such as [8], [9] have shown that machine learning algorithms are accurate for decision making, cost effective and are quick and powerful for data processing. Several studies such as [12] reported that deep learning (DL) techniques, a subset of machine learning characterised by several number of layers in a neural network are suitable to even more complex problems and are capable of handling diverse and unstructured datasets. For example, Muniasamy et al. [12] showed that while conventional ML techniques require handcrafted feature engineering for model development, deep learning models carry out feature engineering automatically.

Several Deep learning (DL) techniques such as convolutional neural networks (CNNs), deep neural networks (DNNs) and recurrent neural networks (RNNs) have achieved impressive results in predictive modelling in many areas including medical research [12]. For instance, Pham et al. [13] created a deep learning framework called DeepCare that learns patterns of several illnesses such as diabetes to predict their future outcomes. Google's DeepMind Health team also used RNNs to predict the onset of acute kidney injury in patients [14]. Dutta et al. [15] used CNNs to predict the occurrence of coronary heart disease in individuals. Thus, it can be inferred that that both conventional ML and DL techniques have the potential to provide medical practitioners new tools and novel ways to better manage their practices. While some studies [8], [16] argue that DL techniques perform better than conventional ML methods, other studies [17], [18] suggest that conventional ML methods produce similar results depending on the type and number of datasets available for training tasks. However, availability of data is usually a challenge for most machine learning studies [19], [20]. Worse can be said about the accessibility of medical related datasets due to its sensitive and controlled nature [19]. Thus, the inaccessibility of data adds to the difficulty of model comparison, accuracy, and the advancement of machine Learning as a whole [21], [22]. This issue can be addressed by adopting data augmentation techniques such as window slicing, image cropping, and the use of generative models to generate artificial data [21], [22] .

This study explores climate-based and diarrhoea-based time series datasets from the nine South African provinces to derive models for diarrhoea outbreak prediction. To augment the data we have available, we will use generative adversarial networks (GANs) to generate synthetic data. In addition, the ML techniques we will use to develop the prediction model are long short-term memory networks (LSTM) because of its ability to

work well with sequential data [16], CNNs because it is a current state of the art in deep learning research [12], [15] and SVMs, a traditional supervised ML method due to its vast success in many prediction studies [8], [10], [17] .

1.1. Motivation and Problem Statement

Diarrhoea related illnesses are prevalent and a leading cause of morbidity in South Africa [23]. In the year 2000, four percent of the total death records among individuals of all ages in South Africa were attributed to diarrhoea [23]. In 2010 and 2015, diarrhoea was reported to be among the top ten leading cause of years of life lost among South African residents [23]. Recently, South Africa witnessed an increase in the rate of diarrhoea [24]. For example, in 2015-2016 provinces like Gauteng, Eastern Cape, Northern Cape, North West and Western Cape experienced an increase in the rate of diarrhoea reported cases, with the highest observed percentage increase of 4 percent in North West [24]. For most provinces, diarrhoea is mostly attributed to nosocomial infections or community acquired resulting from contaminated food and water caused by a range of pathogens [25]. However, other studies [5], [6] show that climate factors and weather variability influence the level of abundance and seasonality of the pathogens present in the environment thus, the prevalence of diarrhoea can be linked to extremities from weather events.

In South Africa, high number of diarrhoea cases caused by bacterial pathogens are reported in the summer months and rotavirus pathogens cases are reported in the winter months [26]. In Western Cape, the warm and dry period from November to May is noted as diarrhoea peak season as it coincides with an increase in the number of reported cases [27]. In addition, studies suggest that the warmer weather worsens the spread of germs and the surge can also be attributed to the severe drought in the region whose occurrences was exacerbated by climate change [27]. South Africa is a climate hotspot and thus will experience an increased frequency and magnitudes of extreme events such as drought, dry spells, heat waves, hailstorms, and veld fires [28]. These events are reported to have the potential to increase water-borne diseases such as diarrheal diseases [4], [5]. Climate factors play a vital role in the long-term trends of infectious diseases such as diarrhoea related ones [5], [6]. The development of a model with the ability to capture complex relationships and long-term dependencies between climate features and diarrhoea may be effective for diarrhoea predictive analysis. Most of the current diarrhoea predictive models although proven useful, are often limited to their reliance on statistical models whose predictive abilities often depend on the assumption of linear relationships or built-in parameter time lags [8], [16].

Machine learning algorithms on the other hand are known for their ability to build and model complex predictive problems and handle high-dimensional data [8], [16]. Studies [8], [29] have also shown that ML methods are good at predicting and diagnosing climatic impacts on public health. For instance, they have been used in West Africa to model the effects of weather and climate on malaria [29]. Hence, they could also be effective in modelling weather and climate impacts on diarrhoea for future projections. However, little work has been done with ML on this application, especially with a focus on case-studies in Southern Africa countries like South Africa. Filling this gap, will make it possible to predict outbreak and vulnerable periods. Furthermore, ML algorithms like CNNs are popular for their powerful feature extraction capabilities [15], LSTM are also known for their ability to capture long term dependencies [13], [16] which may be crucial for time series data modelling therefore, they will be used in this study. Due to the non-linear nature of time series datasets, SVM "a traditional ML algorithm" will also be adopted since they are widely accepted for their ability to solve nonlinear regression estimation problems [30]. Despite the capabilities possessed by these algorithms, there is little empirical evidence about their efficacy for diarrhoea outbreak prediction tasks. Thus, the present study applied them (CNNs, LSTMs and SVMs) in building predictive models that uses time series climate features to predict future number of diarrhoea cases. This information could be useful in reducing the incidence of diarrhoea in the region, which will in turn ease the pressure on the facilities which are already stretched in the health sector.

In an attempt to build ML methods that solve real world problems, relying only on accuracy and error metrics in measuring performance to justify the use of a model might not be enough. Many studies [18], [31], [32] have recommended providing extra measures of confidence to evaluate the performance of a model as a guide, before deciding whether the model is appropriate for a given problem. One example is to compare the performance of a model against others [8], [16]-[18]. In this study, we compare the performance in terms of accuracy of the three proposed models against each other to ascertain which of them is most suitable for diarrhoea outbreak prediction task. Since the performance of an depends on several factors such as its parameter settings and the amount of available training data [33] [18], we augmented the available data with synthetic data which we generated using generative adversarial networks (GANs). We chose GANs because they are popular for their effective capability to generate different types of realistic data [19], [22]. In addition, since there is no default setting for the parameters of an algorithm to guarantee optimum performance, we used relevance estimation and value calibration (REVAC), an evolutionary algorithm in tuning the parameters of the three models because studies like [34], [35] have shown it to provide a good search space for optimum parameter values . Using a single hyper-parameter strategy may also reduce performance estimation bias when all three models are compared with one another.

1.2. Research Objectives

The aim of this research is to use climate variables to forecast the possible number of diarrhoea cases in geographic locations with climate similar to that of Southern Africa, where South Africa will be used as a case-study in this thesis. The chosen case study comprises of several climate variables. However, for this study, we would consider the following: Maximum Temperature, Minimum Temperature, Mean Temperature, Precipitation Rate, Potential Evaporation Rate, Specific Humidity, Surface Pressure and Windspeed because they are the most widely used in climate impact studies [6], [8], [36] . Furthermore, some studies show that they are the main indicators of a changing climate in any geographical location [6], [8], [36].

In this study, we used the listed climate variables as input to predict a range of diarrhoea cases dataset obtained from a clinical source (that is, real data) and the generative model (that is, synthetic data) for each South African province. The main objective of this study is to detect which supervised machine learning techniques (CNN, LSTM and SVM) performs best in terms of high accuracy when predicting the number of diarrhoea cases given a range of datasets (for example, varying proportions of real and synthetic climate variables and diarrhoea datasets) for training and testing. To further address this objective, the following sub-objectives have been formulated:

1. Test the performance of existing deep learning methods such as CNNs, LSTMs and an existing conventional ML method like the SVMs across a range of datasets (that is, varying proportions of real and synthetic climate variables and diarrhoea-based datasets at different testing and training intervals).

2. Investigate the effect of the augmented (a combination of real and synthetic datasets) training and testing data on model performance in terms of prediction accuracy of the three models.

3. Investigate to what extent REVAC parameter tuning can improve the accuracy of the three models.

1.3. Contributions of study

In spite of the fact that technology has evolved, many communities still face lots of challenges in controlling infectious disease outbreaks such as diarrhoea related ones [1], [3], [4] . Currently, little has been done in developing automated diarrhoea outbreak early warning systems, and predicting outbreaks often depends on ad-hoc advice of medical

experts in most communities [37]. This study aims to fill this gap by demonstrating the value of an ML based diarrhoea prediction model that could be used as an automated early warning system for diarrhoea outbreak prediction. The ML model could be extended to predict the outbreak of other infectious diseases, thus a potential contribution to the larger field of disease control.

Another key contribution by the proposed research is the exploration and insight to which of the current ML methods is most suitable for the diarrhoea outbreak prediction task in this study. The anticipated outcome indicates to what extent the ML methods in section 1.2. can be utilized in making diarrhoea outbreak predictions with the datasets available for the proposed case study.

This study also gives an insight as to whether the use of REVAC evolutionary algorithm [34], [35] as a parameter tuning method can improve model performance. It also gives us a deeper understanding on how the amount of data used for training a model can affect the performance of the machine learning model. Furthermore, the combination of the datasets (climate-based and diarrhoea based) used for this study could further strengthen the claim that states that climate factors affect diarrhoea [5], [6], [26], [36].

1.4. Thesis Outline

The rest of the thesis is structured as follows

Chapter 2
This chapter provides the background of this research and introduces the effects of climate factors on diarrhoea, the gap in diarrhoea disease prediction study, different machine learning methods, its applications in disease modelling and previous studies on model performance improvement.

Chapter 3
This chapter describes the specific implementations that were used for this research including details of the machine learning algorithms. It also describes the implementation of the REVAC tuning we used to optimize the ML parameters.

Chapter 4
This chapter describes the datasets, the experiments and details of the synthetic data generation and parameter tuning for all models. It also describes the performance evaluation functions defined for the ML methods, the synthetic data generation method and the REVAC tuning.

Chapter 5
This chapter describes the results of each experiment and assesses the performance of each ML method, the effect of synthetic data in training and testing as well as the parameters we used for REVAC tuning. Results are then presented with visualizations and statistical tests.

Chapter 6
This chapter presents a detailed analysis of the results and relates them back to the initial hypothesis.

Chapter 7
This chapter discusses the conclusions, contributions and limitations of this study and suggests future work.

Chapter 2

2. Background

This chapter provides some background on the global burden of diarrhoea and the impact of climate factors on its severity. It also reviews existing literature on current methods for mitigating diarrhoea, as well as an overview of studies that used climate factors to develop intervention models for diarrhoea outbreak. The applications of machine learning in controlling infectious diseases were also reviewed in this chapter. The focus of this study is South Africa thus more attention will be given to the Southern African region.

2.1. Global Burden of Diarrhoea

The global burden of diarrhoea is widely documented in literature. For example, [3], [4], [6] reported that diarrhoea is one of the worldwide leading causes of death and individual years of life lost. Each year, an estimate of 2 billion episodes and 1.6 million deaths are recorded on a global scale [4], [38]. Another study conducted by Troeger et al. [38] showed that, diarrhoea is the eighth leading cause of mortality among individuals of all ages and the fifth leading causes of death among children under five years of age. In 2016, the diarrhoea death rate in adults tripled, that of children under the age of five and of all the deaths recorded, estimates suggest that the cases were more severe in developing nations [38]. For instance, Troeger et al. [38] reported that Sub Saharan Africa (SSA) alone had an estimate of 1 billion severe diarrhoea episodes with over six hundred thousand deaths, while South East Asia recorded over 470 million cases with a total death rate of over seventy thousand.

In the Southern Africa region, Lesotho, Botswana, and South Africa accounted for the highest diarrhoea case fatality rate [38]. The region recorded an estimate of 80 million episodes with over 35,000 deaths in 2016 [38]. In South Africa, diarrhoea accounts for 3% of the total death records in across all ages, making it the eight most preeminent cause of death in the country [39] . In the year 2000, 8.8% of the total years of healthy life lost for South African residents was also attributed to diarrhoea [26]. Some studies suggest that, its prevalence in the region can be as a result of lack of proper hygiene, poverty, other health conditions including varying weather conditions [26], [27], [39]. Others [6], [8], [26] have also showed that extreme weather events and climate variations affect the rate of diarrhoea infections in a specific location. Therefore, it is important to model the impacts of weather and climate on diarrhoea incidence.

2.2. Impacts of Climate Factors on Diarrhoea

Many studies have shown that climate factors have massive impact on the prevalence of infectious diseases such as diarrhoea [5], [6] . For instance, Alexander et al. [40] explained that variability in climate factors such as temperature, rainfall, relative humidity, and air pressure will be one of the major challenges for developing countries to control diarrhoea. Several observations by [5], [6], [28] have also shown that extreme weather events ranging from heat, cold, drought or heavy rainfall lead to changes in water, food, air quality and the ecology of infectious diseases, and all these pose a threat to humans through increased mortality and morbidity. According to Awotiwon et al. [26], diarrhoea cases are related to changes in temperature and precipitation. Musengimana et al. [27] also showed that for every 1-degree Celsius increase in temperature, diarrhoea cases increased by 8 percent.

Studies have reported that the prevalence of diarrhoea according to seasons and climate conditions varies according to geographic locations [38], [41]. Chang et al. [8], [36] showed that a rise in the incidence of diarrhoea in some Asian countries can be associated with increased rainfall and temperature. For example, Wang et al. [8] showed that diarrhoea in Shanghai, China occurs frequently in the summer and autumn years. Chou et al. [36] also reported that in Taiwan, maximum temperature and extreme rainfall days strongly influence diarrhoea incidence; this can be the effect of runoff due to heavy rainfall which causes contamination in drinking water distribution systems [39], [42]. However, in East African countries like Ethiopia, diarrhoea high risk period usually occurs at the beginning of the dry season [42]. In Southern African region, climate change is projected to lead to warmer temperatures and warming is expected to have a two degree celsius increase compared to the approximate global rising rate [43]; this could lead to an increase in the replication rate of some pathogens, and this in turn can lead to an increase in the incidence of diarrhoea [26], [41]. In South Africa, high number of cases due to bacteria pathogens are recorded in the summer months and high number of rotavirus cases are recorded during winter [26] . In addition, Musengimana et al. [27] reported that the warm months between November and May in Western Cape, South Africa have the highest number of diarrhoea related hospitalizations. However, in Botswana, a peak in diarrhoea incidence usually occurs in wet and dry months of March and October [40].

All these observations indicate that variability in climate factors significantly affect the increase in the rate of diarrhoea incidence, therefore efforts made in understanding the impact of climate change on diarrhoea patterns is critical in controlling its spread. Climate information could also be useful in the development of systems that aid in reducing the spread of diarrhoea.

2.3. Current Methods of Diarrhoea Outbreak Research

This section explores what is currently being done to alleviate the burden of diarrhoea. It explains how medical professionals detect diarrhoea outbreak and it also describes studies that develop models to simulate how changing climate conditions affect diarrhoea.

2.3.1. Human Experts Outbreak Detection Methods of Diarrhoea

The prediction methods of diarrhoea outbreaks by human experts have been sparingly documented in literature. According to Awotiwon et al. [26], [37], outbreaks of infectious diseases including diarrhoea can be defined as the number of cases that is in excess of what would be commonly seen in that season or location. Manatsathit et al. [2] reported that possible outbreak of diarrhoea can be identified when there is an increased rate of hospitalizations due to the disease. In addition, Njidda et al. [44] showed that a high chance of an outbreak can be predicted if the occurrence of at least one confirmed case of diarrhoea is detected from an area that has been identified to be a hotspot or endemic. Elimian et al. [37] & Njidda et al. [44] also identified a diarrhoea outbreak to be when there is detection of the disease from the same area within one week in clusters of persons aged two years or above. Elimian et al. [37] further reported that repeated cases are usually followed by clinical investigations and when a case of diarrhoea is confirmed, an outbreak will be declared.

These studies indicate that outbreaks are detected only when several cases have been reported and confirmed. However, in many African countries, remoteness to health facilities is an issue for potential patients, thus limiting outbreak investigation capacity despite diarrhoea frequency of occurrence [45]. Therefore, there is a need to develop systems to strengthen and improve the surveillance methods that are already in place.

2.3.2. Models for Diarrhoea Outbreak Studies

In recent times, several studies have used computer algorithms to develop models for investigating diarrhoea outbreak in various communities. For example, Chou et al. [36] used the Poisson regression model to predict and quantify the relationship between climate factors and diarrhoea associated morbidity in Taiwan. Constantin et al. [46] used both generalized linear model and poisson regression model to estimate the temporal pattern of diarrhoea by considering environmental factors such as temperature and rainfall. Lloyd et al. [47] used the log-linear regression method to assess the association between temperature, rainfall, and diarrhoea incidence in children under the age of five across the globe. Dhimal et al. [6] used Time-series log linear regression and negative

binomial regression to assess the impact of long-term climate change on diarrhoea epidemics. Yan et al. [48] also used the influence of meteorological variables to develop an autoregressive integrated moving average model (ARIMA) that predicts the daily incidence of diarrhoea in Beijing. In addition, McCormick et al. [49] used the Spatial panel regression model to explore the spatial and temporal patterns of seasonal diarrhoea morbidity in Thailand.

Diarrhoea research models have also been designed for some parts of Sub-Saharan Africa. Azage et al. [42] used the Space-time permutation scan statistics and a negative binomial regression analysis to identify high risk periods and the relationship between climate variables and diarrhoea cases in Ethiopia. Alexander et al. [40] used the autoregressive analysis of covariance model (ANCOVA) and climate factors such as vapour pressure as predictors to analyse the monthly outbreak of diarrhoea in Botswana. In South Africa, Musengimana et al. [27] used the poisson regression model to assess the relationship between diarrhoea cases and temperature variability in Cape Town. In addition, Elimian et al. [37] used basic exploratory data analyses such as histograms and frequency tables to describe the severity of acute watery diarrhoea outbreak in Nigeria.

The findings of these studies though proven useful are based purely on statistical models. Meanwhile, studies such as [8], [16], [30] have shown that traditional statistical models and frameworks are often limited for the analysis of high dimensional, imbalanced, and non-linear data. In addition, these studies [8], [9], [16] explained that the limitations of statistical models can be addressed using machine learning methods. Machine learning models are known to accurately perform statistical data analysis such as classification and regression [9], [30] . Thus, in the present study we used machine Learning techniques to model the influence of climate variables on diarrhoea outbreak.

2.4. Machine Learning

Machine learning (ML) is a multi-disciplinary field that draws concept from various subjects such as artificial intelligence, statistics and biology [50]. It involves the construction of computer programs that can automatically improve performance on a specific task by learning from data [51]. ML techniques can discover knowledge or hidden patterns from large data to make decisions and predictions. They work well for complex problems and can adapt to new data or changing conditions [51]. Other benefits of ML techniques include accuracy, cost effective solutions, quick and powerful processing [9]. ML techniques have thrived in solving real world applications in many fields such as health, environment and finance [20], [51].

Learning by an ML system can be classified based on the type of supervision they receive during training [20]. These are Supervised Learning, An unsupervised Learning, Semi-Supervised Learning and Reinforcement Learning [20], [50], [51]. A supervised ML model is trained using labelled data which provides the algorithm with feedback to evaluate its

training accuracy. Unsupervised learning model on the other hand, gain useful insights on its own from unlabelled data. Semi-supervised learning models are usually provided with input data where a small portion of it is labelled and the rest unlabelled thus, they sit between both supervised and unsupervised learning. Reinforcement learning models do not require labelled input data as well rather, it describes a learning problem where a learning agent must take actions to accomplish a goal in a specific environment to maximize a reward function. In this study, we focused on the supervised learning approach therefore it will be discussed in detail in the following section.

2.4.1. Supervised Learning

A supervised learning model learns a mapping function by observing some labelled input-output pairs (training data) during training [50]. Such labels are usually the desired solution of the task [20]. When the desired output for a learning task is one of a finite set of values, it becomes a classification problem but when the desired output is numerical or continuous, it is called a regression problem [50]. The objective of a regression problem is to find an approximate mapping function that is as accurate as possible, because the chance of finding the exact value for an input-output pair is zero [50]. The workflow of a supervised learning algorithm is shown in Fig. 2.1. Some popular supervised learning algorithms are random forests, support vector machines, artificial neural networks, k-nearest neighbours, decision trees etc. However, in the present study, we used support vector machines and different categories of artificial neural networks to solve a regression task, that is, predicting the possible number of daily diarrhoea cases given some training data. The machine learning algorithms applied in this study are briefly described as follows:

2.4.1.1. Support Vector Machines (SVMs)

SVMs are mathematical models designed based on statistical learning theory and were first proposed by Vladimir Vapnik and Corina Cortes in 1995 [30]. It is a state-of-the-art ML technique that can be used for both linear and non-linear classification and regression [20], [50]. It has been used in applications such as pattern recognition, object classification, and various time series forecasting tasks [30]. SVMs have also been widely adopted in the field of medical research. For instance, Yu et al. [52] used the SVM to detect the presence of diabetes in individuals.

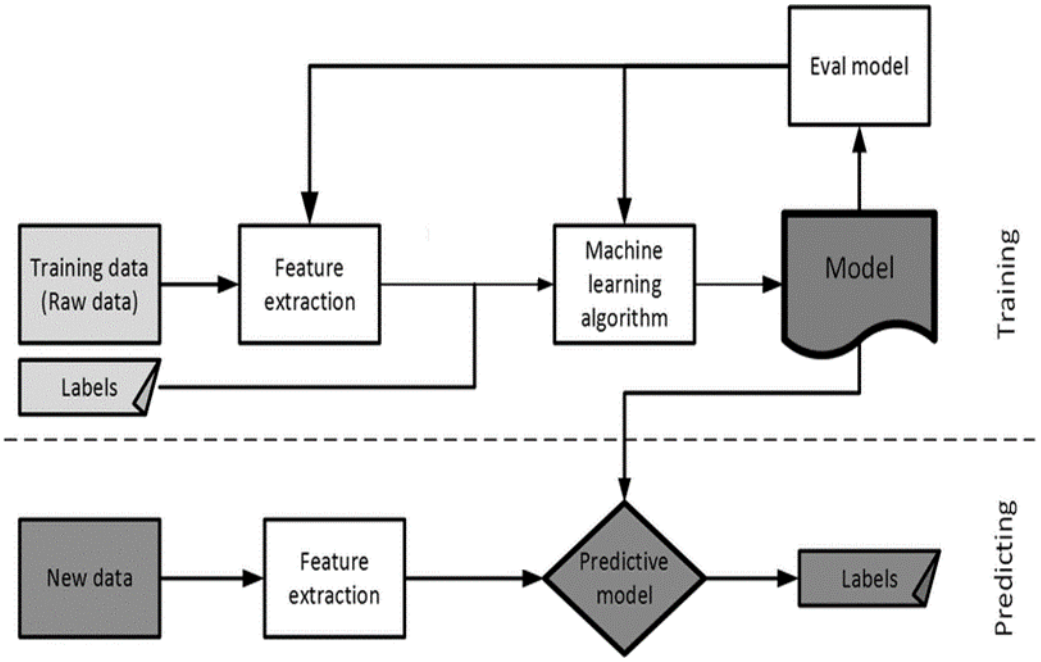

Figure 2. 1: *Supervised Learning Prediction Task Workflow (Source: [55]) The model (designed based on an ML algorithm) takes in some labelled training data. The performance of the model is measured based on its ability to correctly identify the labels. Learning improves by an iterative evaluation and penalization of the model's performance. After a specified training period, the model is given new/unseen data to make predictions based on what it has learnt previously*

The SVM model was able to achieve an accuracy of over 83% when distinguishing between person with and without diabetes. Son et al. [53] also used the SVM to identify predictors of medication adherence in heart failure patients. Even though the sample size they used in training the model was small, SVM was still able to achieve an accuracy of 78%. Although training SVMs can be computationally expensive [30], their advantages are widely documented in literature. Sapankevych and Sankar [30] reported that SVMs are resistant to overfitting and also have the ability to generalize well. Stuart and Peter [50] stated that their non-parametric nature enables them to represent complex and non-linear functions easily. Most SVMs are designed to work with very few parameters thus, the process of tuning its parameters to find an optimal solution may be computationally cheap [30]. Another study by Kilimci et al. [54] discussed the SVM to be capable of providing a description of the learned model depending on the kernel function used during training. In this study we used an SVM to predict daily diarrhoea cases in South Africa. Its applications in the area of infectious diseases is discussed in Chapter 2.5.

2.4.1.2. Artificial Neural Networks and Deep Learning

Artificial Neural Networks (ANNs) are mathematical models which were inspired by the biological learning system of the brain [20], [50], [51]. Several studies such as [8], [9], [11] have shown that they are among the most effective algorithms for modelling complex real-world relationships. In addition, their versatile nature makes them suitable for constructing clustering applications, classification, and regression models [51]. Similar to the human brain, ANNs are composed of several nodes interconnected by links where each node takes several real valued inputs and produces a single real valued output [50], [51]. Each link has a numeric weight associated with it and learning usually takes place by updating the weights.

Brabazon et al. [56] reported that ANNs can be used for both supervised and unsupervised learning. In supervised learning, the ANN is given a set of input-output pairs (training data) over several iterations to find a matching function that minimizes an error. Several kinds of ANN structures have been implemented for supervised learning in literature and each of them has specific computational properties that make them suitable for a specific task. For example, Sharma et al. [10] used a multi layered perceptron with non-directional links and no cycles to predict outbreak of malaria in India while Pham et al. [13] used a recurrent network, whose links form arbitrary topologies to predict a patient's risk of mental illness and diabetes.

An example of a fully connected multi layered perceptron (MLP) is shown in Figure 2.2. MLPs are also a class of feedforward networks [50], [51]. They consist of at least three layers of nodes, an input, an output, and a hidden layer. MLPs often use the backpropagation algorithm, a supervised learning approach for training [50], [51]. Although ANNs are black boxes, studies like [20], [50], [51], [57] have shown that they are very good at generalization and can approximate any function regardless of how complex the problem may seem. They are also versatile and can work well with noisy data [8], [11]. ANNs also form the basis of deep learning [20].

Deep learning (DL) algorithms are an extension of the traditional ANNs. Unlike the traditional ANNs, they use more hidden layers to learn complex patterns in large amounts of data [12]. Several studies such as [13], [58], [59] have shown that they perform better than other machine learning algorithms in many tasks such as image recognition, speech recognition and prediction of drug combination. Another major advantage of deep learning algorithms is they require very little feature engineering by hand unlike other traditional ML techniques [12] [60] . Due to the recent acceptance and applicability of DL algorithms in classification and prediction tasks [12], [20], the present study focused on some DL algorithms such as convolutional neural networks and recurrent neural networks. These algorithms are briefly described as follows:

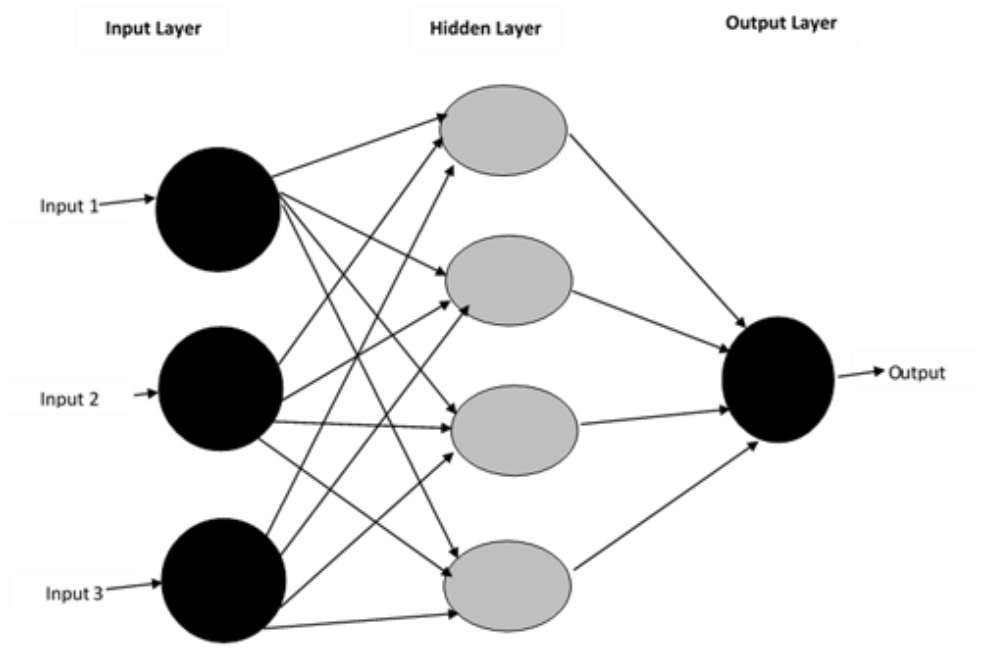

Figure 2. 2: *A Fully connected Feedforward Multi-layer perceptron (Adapted from: [20], [50]).*

2.4.1.2.1. Convolutional Neural Networks

Convolutional neural networks (CNNs) are state of the art deep learning algorithms that have achieved ground-breaking success in many tasks such as image classification and video processing [60], [61]. They have also been used in developing many useful applications for medical diagnosis, object detection, and facial recognition [20], [60]. The use of CNNs are not limited to visual perception, they are also reported to be successful at other tasks such as speech and audio processing, health monitoring applications, and time series forecasting problems [60], [61]. For example, CNNs were used to design an ECG monitoring system to detect abnormal heartbeat of an individual [61]. Huang and Kuo [62] also conducted a comparative study with CNNs and other ML algorithms to make forecasts on the output power of solar photovoltaic energy systems. They [62] found that the CNN model was able to significantly outperform the other ML models because of its robustness and generalization ability.

CNNs process data in the form of arrays, 3D arrays for videos, 2D arrays for images or audio and 1D arrays for signals and other forms of sequence data [60], [61]. The flexible structure of CNNs have been designed to handle any of these data forms effectively. In terms of computation and hardware costs, 1D CNNs have an advantage over 2 and 3D

CNNs because their processing involves only 1D convolutions [61]. In the present study, we used 1D CNNs to predict daily diarrhoea cases in South Africa. The application of CNNs in the area of infectious diseases is discussed in Chapter 2.5.

2.4.1.2.2. Recurrent Neural Networks

Recurrent neural networks (RNNs) are a class of artificial neural networks with feedback connections [16]. They operate by using these feedback connections to store the state of current input events in form of activations [20], [60]. Due to the ability of RNNs to store information, the actions of hidden neurons and output neurons might be determined not just by the current inputs and activations in the previous layers, but also by inputs and activations at earlier times [20], [60]. It is a state-of-the-art algorithm for sequential tasks such as speech and text recognition [60]. Although conventional RNNs have proven useful in many applications, they do not capture long-term dependencies during training due to vanishing and exploding gradients [13],[60].

Long-short term memory networks (LSTMs) a special kind of RNN were formulated to address the issue of vanishing and exploding gradients [13],[60]. They are commonly used to handle sequential tasks such as time series forecasting [16]. In addition, they have been successfully implemented in biomedical research, speech recognition and language modelling tasks [13], [20]. For instance, Pham et al. [13] conducted a study with an LSTM, that uses historical medical data to predict the future mental state of an individual. Helmini et al. [63] performed an investigation comparing the performance of LSTM, Random forests and Extreme gradient boosting and their applicability to forecast sales based on historical sales records. They [63] found that the LSTM model was able to outperform the other models due to its ability to preserve information and identify temporal relationships within the training data. In the present study, we used LSTMs to predict daily diarrhoea cases in the nine South African provinces. The application of LSTM for modelling infectious diseases is discussed in Chapter 2.5.

2.4.2. Machine Learning Applications for Infectious Diseases

The use of machine learning as a tool for medical analytics has become increasing popular in the last few decades [9]. ML has been applied to several aspects of medicine and public health, ranging from applications for computer vision, genomics, disease diagnostics, drug discovery, outbreak detection of infectious diseases, among others [9], [12]. The severity of infectious diseases has made many researchers focus on applications that aid in reducing their widespread occurrence to avoid outbreaks and epidemics. For instance, [10] used support vector machines (SVMs) and artificial neural networks (ANNs) to predict the outbreak of malaria in India, the model was able to give at least two-weeks

lead time for authorities to intervene and minimize risks that could arise from a pandemic. They used climate factors as input variables to classify the possibility of an outbreak. After training and testing, both models were successful, although the SVM model outperformed the ANNs by over 12% in terms of accuracy. Adamker et al. [64] also used SVM and ANN classifiers to predict the chances of hospitalizations due to Shigella and the shigella specie responsible for those hospitalizations. The Shigella clinical records used contained information such as age, shigella specie, year, month, among others. They used 30% of the dataset to evaluate the algorithms' accuracy, and both SVM and ANNs had an accuracy of over 92% for both tasks. The high accuracy of both models indicate that these ML models could be used to strengthen treatment of disease caused by Shigella. Akbar et al. [65] used SVMs and AdaBoost ensemble model to develop a hybrid model that accurately detects the presence of Hepatitis in individuals. Successful and early detection of the virus could reduce the risk of death an individual. In [65] a SVM was used to select features for the AdaBoost model and they found that the SVM model was able to improve the prediction accuracy of the AdaBoost model by 6.39%.

Some studies have applied deep learning models for infectious disease research. These studies also show that some deep learning models were superior in task performance when compared to other ML models. In China, Jia et al. [16] carried out a comparative study between a Gradient boosting model and an LSTM network. Both models were used to successfully predict the outbreak of several infectious diseases such as Typhoid and Malaria. For all predictive tests carried out, the LSTM network used only three input features and was able to outperform the gradient boosting model even though it trained with eleven input features. Chae et al. [66] also used DNN and LSTMs with temperature, humidity, and social media data to successfully predict the outbreak of infectious diseases such as Chicken pox and Scarlet fever in Korea. Both models were compared, and the accuracy of LSTM model was 5% higher than the DNN. The study also reported that both models could help to strengthen the current public health surveillance system in the country. Abideen et al. [67] assessed the performance of a hybrid bayesian-convolutional neural network on two tuberculosis (TB) benchmark datasets, to detect the presence or absence of TB in an individual. The accuracy of the model was compared to some classical CNN algorithms such as ResNet, AlexNet, and VGG19 network. Observations showed that all the CNN based models used had an accuracy of at least 70%. Fuhad et al. [68] conducted a comparative study with CNNs, SVMs and k- nearest neighbours to detect the presence of malaria parasites from microscopic images. The models were able to automate the manual process of malaria detection by clinicians. The CNN model outperformed the other models by obtaining an accuracy of 99.23%.

All these studies indicate that the application ML algorithms especially SVMs, CNNs and LSTMs has been successful in various aspects of controlling infectious diseases. However, very little has been done with ML with regards to diarrhoea outbreak control, specifically in Africa.

Few studies have adopted ML techniques for diarrhoea outbreak studies in other parts of the world. For instance, Wang et al. [8] used ANNs to predict the outbreak of infectious diarrhoea in Shanghai province of China. They [8] reported that the ANN model gave one-week lead time prediction information on the possible number of diarrhoea cases for that province. However, the dataset they used for the study composed of 209 weeks which means they used only 209 data points for training and testing the model. Meanwhile studies such as [21], [69] have shown that Neural Network models tend to overfit when a small sample size is used for training. Fang et al. [70] used a Random forest model to predict the outbreak of diarrhoea with climate information in Jiangsu province of China. Although their Random forest model was able to predict weekly outbreak of diarrhoea disease, it was validated with only the Autoregressive integrated moving average (ARIMA) model. However, studies such as [16], [70] have clearly reported that ARIMA models are limited to modelling problems with linear relationships which may not be the case for diarrhoea and climate factors. These diarrhoea outbreak studies although proven useful have some limitations, thus in this study, we used a traditional ML method, "SVM" and two deep learning algorithms, "CNN and LSTM" to predict the outbreak of daily diarrhoea cases in South Africa. These models were chosen because of their comparative advantages in their applications for infectious diseases. In order to make robust conclusions, we used a variety of datasets (real and synthetic datasets) to train and validate each model. A summary of ML applications for infectious diseases can be seen in Table 2.1.

2.4.3. Current ML Methods Being Applied in the Fight Against COVID-19.

The 2019 novel Coronavirus (COVID-19) disease was declared a pandemic on the 11th of March 2020 by the World Health Organization [71]. As of 8 Dec 2020, over 66 million cases and 1.5 million deaths were reported across the world [71]. Its continued spread has made researchers across the globe work tirelessly to better understand it and pursue possible solutions to reduce its transmission and spread. Several organizations and researchers have already been able to launch different platforms, and applications in the fight against COVID-19, in most of which machine learning has played a major role. Current examples of such applications are briefly highlighted as thus:

Senior et al. [72] used ResNets, a convolutional neural network with hundreds of convolutional layers to develop a model called AlphaFold. AlphaFold was able to predict six different structures of proteins related to COVID-19. Once the protein structures of COVID-19 are known, it might be possible to predict which drugs can effectively contain those proteins [73]. Hu et al. [74] trained a multi-task deep neural network to identify a list of COVID-19 protein structures which were subsequently used as potential targets by

the model to predict commercially available drugs that could bind these proteins. Furthermore, Beck et al. [75] used Google's BERT (Bidirectional Encoder Representations from Transformers) framework to develop a deep learning-based drug target interaction model. BERT [76] is a multi-layer bidirectional transformer encoder and a pretrained deep learning framework mainly used for natural language processing tasks. The model was also trained with a wide variety of antiviral drugs and target proteins; it was able to identify commercial drugs that could contain COVID-19 viral protein structures. Other studies such as [77] used naïve bayes algorithm to predict which commercial drug could be used for COVID-19 treatment. The model was trained to classify several labelled drugs and was able to achieve a classification accuracy of about 73%.

The applications mentioned above were attempts to investigate the potential efficacy of existing drugs for the treatment of COVID-19 disease. Some studies have also made attempts to devise novel drug and vaccines against COVID-19. Zhavoronkov et al. [78] created a new drug molecule with an ML-based framework. They used several input features to train 28 different ML methods including Generative Adversarial Networks and genetic algorithms. Each of the models were further optimized using reward functions based on a reinforcement Learning technique. In an attempt to discover vaccine candidates, Ong et al. [79] used Vaxign-ML, a supervised ML framework for vaccine development to predict which viral protein will serve as the best vaccine candidate. Several bacterial and viral protective antigens were fed as input data to different algorithms such as SVM, random forest and XGBoost. After the models were trained and validated, the XGBoost recorded the highest accuracy.

Some researchers have also focused on forecasting COVID-19 cases and deaths. For instance, a CNN was used to predict the daily number of confirmed cases in China [73]. Other models (LSTMs, MLP and gated recurrent units (GRU)) were trained alongside the novel CNN but the CNN outperformed the others with a very high margin. To confirm if predictions made by clinicians were accurate, Bandyopadhyay and Dutta [80] used LSTMs and GRU to evaluate how close their predictions were to actual confirmed cases. A combined LSTM-GRU architecture outperformed both individual models with an accuracy of over 10%.

Other studies such as [73] [81] used ML methods such as LSTMs, GANs and fully connected networks to forecast the risks of an outbreak in several communities. Applications for medical imaging diagnosis that determine if a person has contracted the virus have also been developed with CNNs and ResNets [73]. In addition, CNNs achieved an accuracy of over 98% when it was used to differentiate the SARS-CoV-2 strain from other similar strains [82]. This application could be extended to improve the accuracy and reliability of current COVID-19 diagnostic tests. Although the above findings further prove that ML algorithms are applicable to a wide range of infectious diseases, the new COVID-19 virus remains a pandemic, and more study still need to be done.

2.4.4. Limitations of Machine Learning Algorithms

The limitations of ML algorithms are widely documented in literature. For example, studies such as [21], [69], [83] reported that when the dataset available for training an ML algorithm is small, the algorithm may overfit the training data and may not properly generalize the problem at hand. For example, Yang et al. [84] used an LSTM to predict the next character in a given sequence. They [84] used datasets of two sizes for training and found that the LSTM performed better when the dataset with a larger sample size was used for training. However, studies like [19], [21] have argued that the availability of data is usually a challenge for most research exercises. Worse can be said about the accessibility of medical related datasets due to its sensitive and controlled nature. This issue can be addressed by adopting data augmentation techniques to generate artificial data [19], [21], [22].

Table 2. 1: Summary of ML applications for infectious diseases

Methods	**Contributions**	**References**
SVMs and ANNs	Malaria Outbreak Prediction	[10]
ANNs, SVM	Prediction of Shigellosis outcomes	[64]
SVM, AdaBoost	Hepatitis Disease Detection	[65]
LSTM and Gradient Boosting	Typhoid, Malaria, Cholera Outbreak Prediction	[16]
DNN, LSTM	Malaria, Chicken pox and Scarlet fever Outbreak Prediction	[66]
Bayesian-Convolutional Neural Network	Tuberculosis Disease Detection	[67]
CNNs, SVM, KNNs	Malaria Parasite Detection	[68]
ANNs, SVMs, Random Forests	Diarrhoea Outbreak Prediction	[8]
Random Forests	Diarrhoea Outbreak Prediction	[70]
CNNs, DNNs, Naïve Bayes, Google's BERT.GANs, SVM, Random Forests, XGBoost, LSTM, GRU	COVID-19 Research	[72]-[75], [77]-[82]

One popular method of data augmentation involves the use of generative adversarial networks (GANs), a class of neural networks to generate realistic looking data [19], [22]. Their massive success has aided the advancement of various real-world applications for security, fashion, and video games, among others. However, in recent years, the application of GANs have been extended to other areas of research to aid studies like natural language processing and time series forecasting [19], [22].

For instance, Wen et al. [22] reported that GANs were successful in generating realistic datasets for various research disciplines to aid data augmentation for predictive analysis. In the field of medicine, Esteban et al. [19] used GANs to generate synthetic medical time series data. They [19] used the synthetic data as a training set on a Random forest classifier and the model was able to achieve an accuracy of 97% when it was tested on real world data. GANs were also used to generate synthetic patient records that included information such as their diagnosis and medications [19]. A qualitative evaluation was conducted and findings showed that it was difficult for medical experts to differentiate between the diagnosis and medication recommendations of a real doctor and the diagnosis and medications data generated by the GAN [97]. Che et al. [98] used synthetic time series electronic health records (EHR) data generated by GANs to augment real world time series EHR data. A CNN was used to compare predictions made with the real-world data and predictions with the augmented data. Their results showed that utilization of the augmented data was able to boost the CNN's task performance.

Another limitation of ML technique is that most ML algorithms are designed to work with lots of parameters that have significant control over their behaviour and performance [58]. Manually tweaking these parameters may be difficult and would also require prior and expert knowledge [58], [83]. Since there is no default setting for the parameters of an algorithm to guarantee optimum performance, it is important for one to adopt techniques to tune the parameters of an ML algorithm. One popular method is to use evolutionary algorithms to search for an optimal solution given a wide range of possible parameter values [34], [35]. Karegowda et al. [85] used an evolutionary algorithm to initialize and optimize the structure of a neural network that was used for diagnosing diabetes in individuals.

In this study we used generative adversarial networks (GANs) to generate artificial data for training because of their applicability to a wide range of problems [19], [22]. We also used relevance estimation and value calibration (REVAC), an evolutionary algorithm in tuning the parameters of our ML models because studies like [34], [35] have shown that using evolutionary algorithms such as REVAC to tune parameters improves the performance and accuracy of most algorithms.

2.5. Summary

In recent years, deep learning algorithms such as CNNs, and traditional ML algorithms such as SVMs have been widely used to develop countless important predictive models in the health care field. Despite their successes, their applications are still comparatively limited. For example, CNN has mainly been applied to image processing and classification problems but there is a dearth of literature on its application for diarrhoea outbreak. Although LSTMs has been used to predict the outbreak of some infectious diseases, very

little has been done to predict diarrhoea outbreak. The few ML models that have been used to predict diarrhoea were conducted outside Africa and are often constrained to data availability which greatly affects the performance of most algorithms. This study aims to fill this gap by adopting CNN, LSTM and SVM algorithms to predict the daily number of diarrhoea cases in South Africa. To address the issue of data availability, we boost the size of our training data with synthetic data generated with GANs.

Chapter 3

3. Methods

The focus of this study is to use ML methods to predict the possible number of daily diarrhoea cases in each province and asses the predictive performance of each ML method used. This chapter gives details on each of the ML algorithms used in this research. It also describes how the REVAC hyper parameter tuning algorithm was implemented.

3.1. Convolutional Neural Network Architecture

CNNs are a class of feedforward, deep Neural Networks that consists of multiple convolutional and activation layers, pooling layers, and a fully connected layer [59] as shown in Figure 3.1. In the convolutional layer, filters are applied to the input array in order to identify the features of the input data using the convolution operation. The convolution operation is a mathematical operation used for feature extraction [60] . Even though the size of the input array reduces after the convolution operation is performed, the most important features are still preserved, and the output of a convolution layer is called a feature map [60].

The activation layer is an extension of the convolution layer; here, after every convolution operation, the feature maps are passed through non-linear activation functions such as the Rectified linear unit (ReLU). The activation functions allow CNNs and other neural networks approximate almost any non-linear functions [59]. Other functions that can be used instead of ReLU are tanh or sigmoid but the ReLU is preferred in most situations [59]. After activation functions have been applied, the output feature maps are fed to the pooling layers.

Pooling layers are designed to condense the information on the feature maps by summarizing its parameters [20], [59] . In addition, pooling layers combine semantically similar features together [60]. Examples of pooling functions are Max pooling, Average pooling, and Sum pooling [20], [59]. After iterating through several layers of convolution, activation and pooling, the final output is computed in the fully connected layer of the network [59], [60]. The fully connected layer uses these features to make decisions based

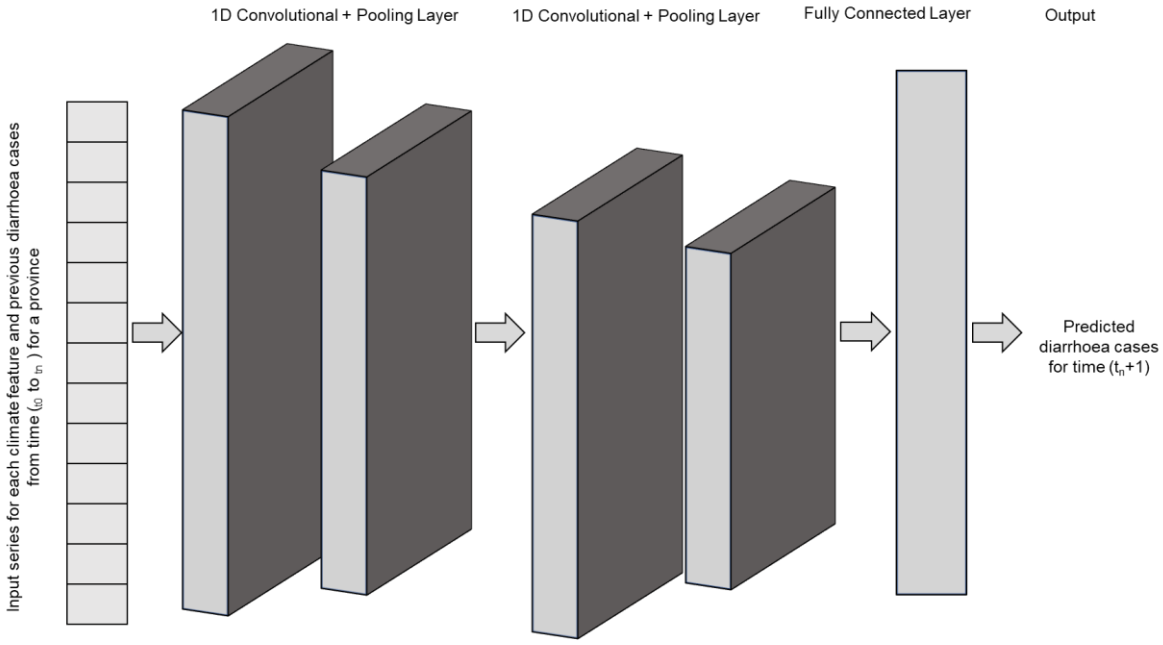

Figure 3. 1: *A Simple CNN Architecture with two convolutional layers. The output of the last pooling layer is fed into a vector of activations and finally into the fully connected layer. The output neuron with the largest activation will be the network's decision/prediction to the problem.*

on the problem specification. Similar to most Neural Networks, CNNs are trained with backpropagation and gradient descent [59], [60].

In this study, our CNN model was designed with 1D convolutions to match the format of our input data which is 1D and sequential in nature. Studies have shown that 1D CNNs are effective for several applications such as time-series forecasting, anomaly detection, text classification and health monitoring [61]. The performance of most neural networks including the CNN depends on its parameters and how they are configured [58]. Some of the important CNN parameters are, the number of convolution layers, filter size, number of epochs for training, etc [58].

Fig 3.1 gives a brief overview of the input data fed into our CNN model during experiments. It also shows the expected output after the input data has passed through a series of convolutions and activations. Our experiment section (see section 4.5.1.2) gives details on the framework we used to design our CNN model, the parameters we chose to tune and the methods we used in tuning those parameters.

3.2. Long-Short Term Memory Network Architecture

LSTMs are another example of Neural Networks under the category of RNNs that addresses the issue of exploding and vanishing gradients [13]. They consist of memory blocks which contain memory cells that allow gradients flow through long sequences. The memory cells maintain its state over time and is managed by gating units that control how the memory cell memorize, erase, and expose information [13], [20]. These gating units which are the input gate, the output gate and the forget gate uses sigmoid functions that set elements of each gates to values in the range between zero and one [13]. The input gates supervises how input activations are added into the memory cell while the output gate supervises which part of the memory cells are read into the rest of the network and finally, the forget gate supervises which part of the memory cells are erased [20].

In this study, our LSTM model was designed to predict future outbreak of diarrhoea cases in each South African province. The performance of the LSTM also depends on how its parameters are configured [58]. Some important LSTM parameters are number of hidden units, optimizer, batch size, number of epochs for training, etc.

Fig 3.2 shows the basic workflow of our LSTM model. The input data fed into our model during experiments are input series for each climate feature and previous diarrhoea cases from time (t_0 t_o t_n). The final output of the model is the predicted number of diarrhoea cases for time t_{n+1}. Our experiment section (see section 4.5.1.2) gives details on the framework that was used to design the LSTM model, the parameters we chose to tune and the methods we used in tuning those parameters.

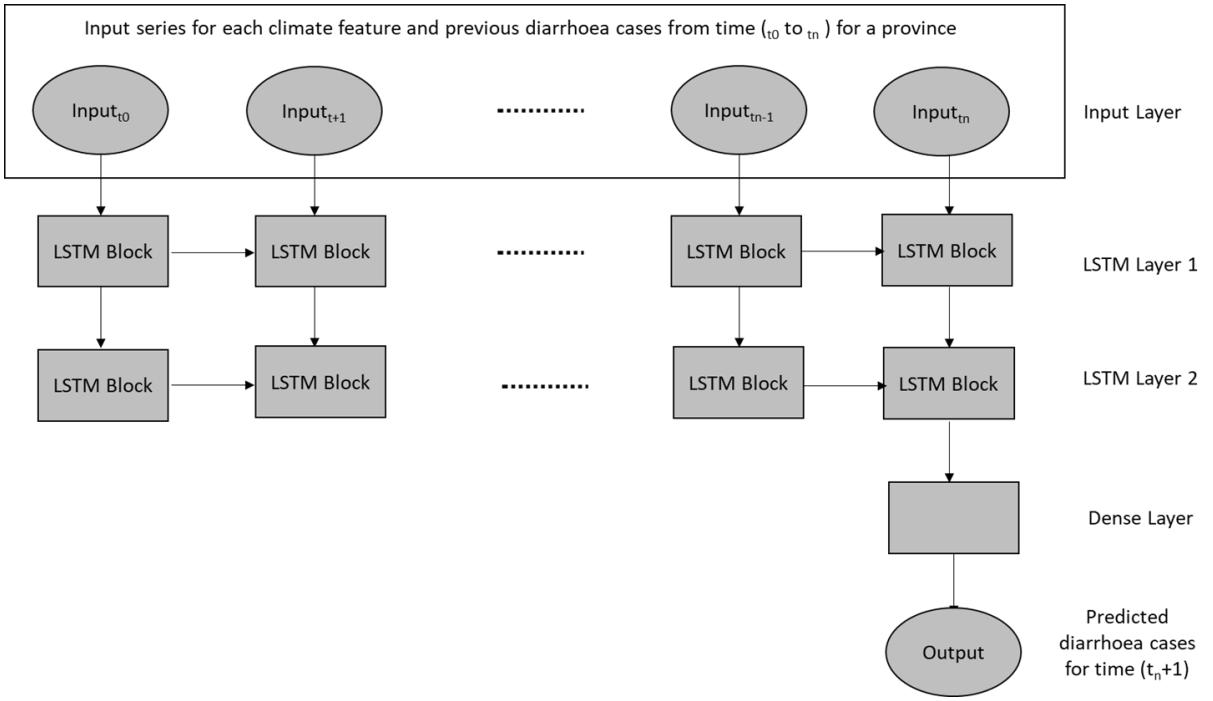

Figure 3. 2: *Basic Structure of our LSTM model with two LSTM layers.*

3.3. Support Vector Machines Architecture

The main function of an SVM is to find hyperplanes capable of creating margins that separates datapoints in a hyperspace[20], [50] as shown in Figure 3.3. An SVM algorithm involves the projection of data points in a training dataset with n input variables into an n-dimensional surface called a hyperplane; the hyperplane that maximizes the distance between the data points (that is, the hyperplane with the maximum margin) will be chosen [50]. The larger the margin, the lower the generalization error [50]. New samples are mapped onto the same space and the binary outcome is predicted based upon which side of the hyperplane each sample falls on. In a situation where the training data is not linearly separable, SVMs uses a technique called the kernel trick to separate the nonlinear data in a higher dimensional space [20], [50]. They use kernel functions to find separators in the high dimensional feature space. Some notable examples are the polynomial, gaussian and RBF kernel functions [20]. Unlike neural networks, SVMs uses fewer parameters that depends on the kernel function chosen during training. In this study we used the RBF kernel function whose main parameters are the regularization parameter, 'C' and gamma. More details on how these parameters were set for all our experiments are given in Section 4.5.1.2.

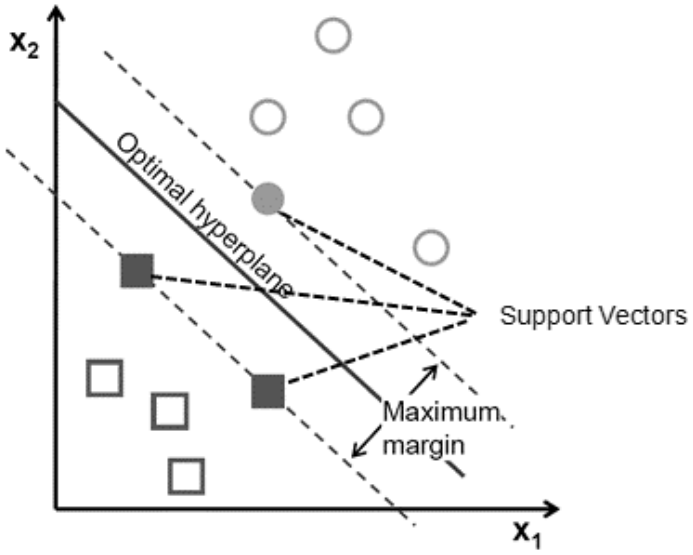

Figure 3. 3: *An SVM trained with samples from two classes (Source: [20], [50]). The data points that fall on the dotted lines are samples from the training dataset that are closest to the decision boundary. They are also called support vectors and determine the margin with which the two classes are separate. Changing or deleting the support vectors will change the position of the hyperplane.*

3.4. Relevance Estimation and Value Calibration

REVAC is an evolutionary method formally designed to tune the parameters of Evolutionary algorithms [34], [35]. Given an objective, a population of parameter vectors and n number of iterations, REVAC explores, selects, and evaluates a set of possible parameter values. By adopting some concepts of evolution, such as mutation, recombination, selection and replacement, it improves and updates the distribution of the parameter vectors such that after each iteration, there is a high chance of obtaining optimal performance when a combination of those parameters values are adopted for training an algorithm [34], [35]. Figure 3.4 gives a brief overview of how REVAC is used to tune the parameters of the ML algorithms. Details of how REVAC algorithm is implemented can be summarized in the following steps:

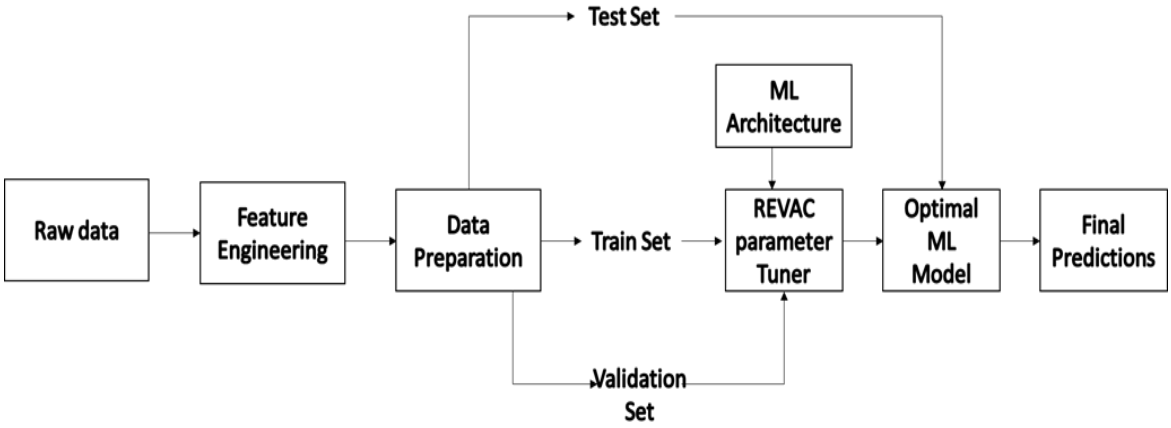

Figure 3. 4: *Workflow of REVAC Parameter Tuning.*

Details of how REVAC algorithm is implemented can be summarized in the following steps:

1. Initialize a population of m parameter vectors and define a utility function (objective).
2. The performance of each new vector is measured based on the utility function.
3. Select n vectors with the highest measured utility to become parents of a new child vector.
4. Create one child by performing recombination with the selected parents. Recombination is a multi-parent crossover operation with uniform scanning [34], [35].
5. Mutate the offspring created from the recombination step. A mutation interval is calculated, and a random value is uniformly chosen from this interval [34], [35].
6. The new offspring replaces the worst performing vectors in the population.
7. The performance of the new vectors are measured with the utility function and all steps are repeated until a stopping condition is met.

In this study, we used REVAC to tune the parameters of three ML algorithms we adopted. See section 4 on details of how REVAC was implemented.

3.5. Generative Adversarial Networks

GANs as shown in Figure 3.5 are a common example of deep generative models that are used in generating realistic high-dimensional objects such as images and sequences. They were formally proposed for image generation in 2014 by Goodfellow et al. [99]. GANs can also be described as a class of ML algorithms that consists of two neural network models called a generator and the discriminator. These models are usually trained via an adversarial process whereby the generative model captures a random distribution and outputs some synthetic data and the discriminative model estimates the probability that

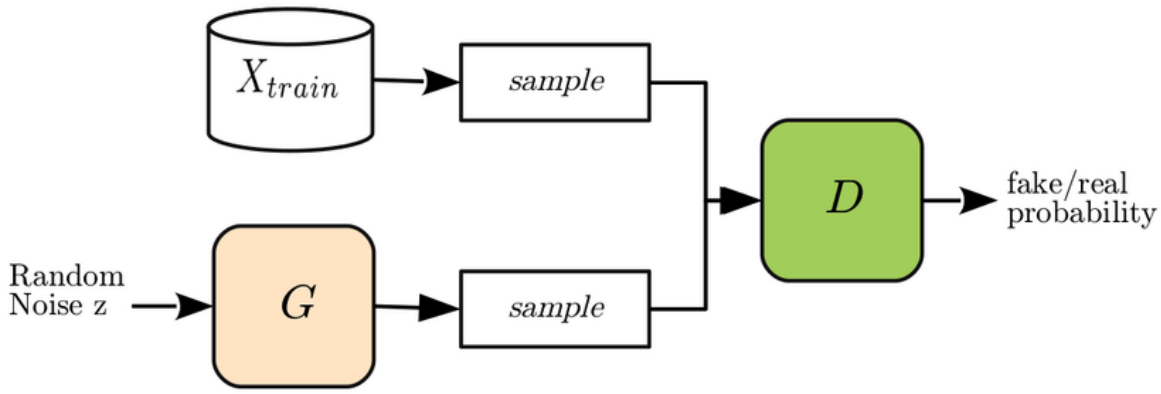

Figure 3.5: *Basic workflow of a GAN (Source: [100]). G and D represents the generator and discriminator respectively while X_{train} represents the inputted training data.*

the synthetic data came from the input training data rather than the generator model [20][99]. When training begins, the discriminator easily tells that the synthetic data is fake. However, as training progresses, the generator gets better at generating realistic samples which easily fools the discriminator. The training procedure is divided into two stages and achieved with backpropagation whereby the aim of the GAN is for the generator to maximize the probability of fooling the discriminator when distinguishing between synthetic and real samples [20]. In the first stage, the discriminator uses backpropagation to optimize its weights and in the second phase, the weights of the discriminator are kept constant while the weights of the generator are affected by the backpropagation algorithm. In this study, we used GANs to generate synthetic samples to augment the available real-world data. See section 4 on details of how GANs was implemented.

3.6. Summary

In this study, we used three ML algorithms (CNNs, LSTMs and SVM) to predict daily number of diarrhoea cases in South Africa. The approach that was used to design our models was similar to what have been done in previous predictive ML studies in the health care domain however, there were some slight peculiarities. For instance, in designing our CNN model, 1D convolutions were used. We used GANs to generate synthetic data for data augmentation. REVAC tuning was also used to tune the parameters of all three ML algorithms. REVAC tuning is an evolutionary strategy that has mainly been applied in optimizing evolutionary algorithms.

Chapter 4

4. Experimental Design

The experiments[1] in this study were designed to determine the most effective ML algorithms in terms of performance accuracy (proposed in section 1.2.) for predicting the possible number of daily diarrhoea cases (that is, to give one day lead time) in each of the 9 provinces separately with respect to a given set of training data. The performance of each of the proposed ML algorithms may be influenced by several factors such as:

- Input data for the ML algorithms (CNN, SVM & LSTM) described in chapter 3.
- Choice of parameters.
- Amount of training and testing data.
- Method of parameter tuning.

The effects of the above factors were investigated through a series of experiments for each ML algorithm. This section aims to give a detailed explanation of the various experiments carried out to show the effects of the factors above.

4.1. Datasets

The datasets used for this study contained 9 different features and can be categorized into two:

a) Health data: It consists of Clicks pharmaceuticals daily sales records of loperamide, an anti-diarrheal compound that has been evaluated in the treatment of patients with chronic non-specific diarrhoea in South Africa and other parts of the world [1] . The data contains a 10- year period of total number of loperamide purchased between November 2008 and March 2018 in every Click pharmacy across each of the nine South African provinces. This data was used as a proxy for diarrhoea cases in the region. In this study, the number of diarrhoea cases per day for a specific province was computed as the number of loperamide sales per day associated with the province. Throughout the experiments, the daily loperamide sales data was referred to as daily diarrhoea cases dataset.

b) Climate data: Climate factors such as Maximum temperature, Minimum temperature, Air temperature, Specific humidity, Potential evaporation rate,

[1] https://github.com/aminalawal/Predicting-Diarrhoea-Outbreak-with-Climate-Change

Precipitation rate, Surface pressure, and Wind velocities for each province between the period of November 2008 and October 2019 were obtained from National Centre for Atmospheric Research (NCAR)/National Centre for Atmospheric Prediction (NCEP) (*https://psl.noaa.gov/*). These data are known as reanalysis datasets. Please see for [86] more information.

4.2. Lag Variables

For most time series prediction studies [66], making forecasts based on past observations is usually recommended because patterns of the past are likely to be repeated in the future. Therefore, the selection of past values (lags) from all input features is a crucial step and may be important for learning during model training. For each experiment, we tested the predictions of the three ML algorithms with respect to four different lag periods from all input features. The lag periods we considered were a lag of 1 day, lag of 5 days, lag of 2 weeks and lag of 3 weeks. For example, a lag of one day means that the predictions made by a model for the 6th of January 2018 was made with input variables (for all features) of the 5th of January 2018 while a lag of 5 days means predictions for the 6th of January 2018 was made with input variables (for all features) of the 1st to the 5th of January 2018.

4.3. Data Pre-Processing and Post Processing

The original climate and diarrhoea cases datasets for each province collected for the study were ordered in the form of time series. The datasets for each province were processed separately since the diarrhoea case prediction model for each ML algorithm was developed per province. The diarrhoea cases data collected was in a daily format while the datasets for the climate features were collected in an hourly format. To use the climate features to predict daily diarrhoea cases, we had to change the format from six-hourly to a daily average format. This was achieved by aggregating the six-hourly data values for every 24 hours and dividing by the mean which is four (since, every six-hours in a 24-hour time period will contain four data values). For each province, the climate features datasets had more data points than the diarrhoea cases datasets but for our experiments, we picked data points based on the same date for both datasets so that the training and testing data for each province will contain an equal number of data points for both datasets. While there was no occurrence of missing values for the climate features datasets the very few occurrences encountered for the diarrhoea cases datasets were handled as no cases for that day. For all experiments and models, the ratio of training to testing data were divided in the 70/30 ratio. The datasets with the earlier dates were used for training while the datasets with later dates were used to test and verify the accuracy of the models.

Table 4. 1: Overview of all Experiments. Experiment I, II & III are fully described in sections 4.5.1, 4.5.2 & 4.5.3 respectively

Experiment Name	Experiment Description	ML Methods Used	Method of parameter tuning	Parameters Tuned	Data Used	Research Objective Addressed by the Experiment
Experiment I	Predictions with original data only	SVM, LSTM, CNN	Grid Search	See Tables (4.2a - c)	Original data only	See objective 1 in Section 1.2
Experiment II	Predictions with original data and synthetic data generated by GANs	SVM, LSTM, CNN	Grid Search	See Tables (4.2a - c)	Original and Synthetic data (see tables 4.3&4.4)	See objectives 1 & 2 in Section 1.2
Experiment III	Predictions with ML methods whose parameters were tuned with REVAC and whose input data were the same as Experiment II	SVM, LSTM, CNN	REVAC tuning	See Tables (4.5a-d)	Original and Synthetic data (see tables 4.3&4.4)	See objectives 1 & 3 in Section 1.2

Another pre-processing step we took before training is the selection of a lag period for all our input features. Once a lag period is selected the previous values based on the selected lag period and training data ratio is processed and fed as input to the prediction model. To improve the efficiency of ML models, normalization techniques are usually adopted to speed up convergence and learning process. Normalization is also required when the features have different ranges/scales. Since all the climate features and diarrhoea cases have different numeric scale, we normalized the values of these features for our experiments. The normalization technique adopted depends on the type of ML algorithm and the type of dataset available for training (numerical/categorical). For all experiments, the normalization technique adopted for our CNNs and LSTMs is the Min-Max Normalization /Scaling technique from the python Scikit-Learn library because it is largely adopted for most neural network regression models. For the SVM model, we adopted the Standard Scaling technique from the python Scikit-Learn library for all experiments since SVMs assume that the data given as input is within a standard range.

In addition, we observed that after normalizing/scaling the input values of our models, the output/result will correspond to the normalized range. Hence to interpret the results from the models, we inversed the normalizing/scaling initially performed to transform the output back to its original scale.

4.4. Performance Evaluation Criteria

To compare and evaluate the performance of an ML algorithm, several evaluation criteria such as Mean absolute error (MAE), Correlation Coefficient (R), Mean absolute percentage error (MAPE), Root mean square error (RMSE) and Coefficient of determination (R2) have been used in previous research [8], [16], [66]. Although proven useful, some of these error metrics such as the MAE make use of absolute values which is often avoided in many mathematical studies, the RMSE on the other hand avoids the use of absolute errors and is superior at disclosing differences in model performance [87] . Other studies such as [88] have shown that a specific metric may be chosen for a certain purpose. For instance, if the absolute values of the estimated predictions by a model is important, the RMSE may not be an appropriate metric. However, if evaluations based on understanding of predictions is desired, the RMSE should be used [88] . In this study, we used the RMSE to evaluate the accuracy of our ML models in all experiments not just because we aim to compare differences in ML model skill but also because it is widely used in many prediction studies including studies for climate research and infectious diseases prediction [8], [16] .

RMSE is the square root of the mean of the squared differences between actual outcomes and the predictions made by a model. It is calculated using the equation below:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_t - y\hat{t})^2} \qquad (4.1)$$

In the equation (4.1), y_t is the actual value while $y\hat{t}$ is the predicted value and n is the total number of observations to be analysed. The model with the smallest RMSE error is considered to be the best performing model in terms of prediction accuracy.

4.5. Experiments to Determine the Best Performing Algorithm

To determine the best performing ML algorithm for predicting the possible number of daily diarrhoea cases (that is, prediction for one day lead time), we compare the RMSE from the predictions made by the three ML algorithms with different factors with respect to the factors listed in section 4. For every algorithm and experiment, lower RMSE errors indicate better prediction accuracy. The experiments designed to investigate the effects of these factors were broken down into 3 sections. They are:

1. Experiment I: predictions with original data only
2. Experiment II: predictions with both original and synthetic data
3. Experiment III: predictions with both original and synthetic data and REVAC parameter tuning

4.5.1. Experiment I: Predictions with Original Data Only

The first set of experiments conducted for this study were implemented with the original data obtained from a clinical source (diarrhoea cases) and 8 climate features data. The experiments were carried out across four different lags periods (see section 4.2) for each ML model across all provinces. After pre-processing, there were 3760 data instances across all 9 features available for both training and testing. This was divided in the ratio 70/30 for training and testing. Each ML algorithm was trained and tested with a specific set of parameters.

4.5.1.1. Grid Search Parameter Tuning

Grid search is a parameter tuning technique that trains an ML algorithm with a combination of possible parameters (specified by the user) on the training set and evaluates and outputs the best parameters based on a given performance metric. We used RMSE as the performance metric for this study. The Grid search parameter tuning was implemented with the python Scikit-Learn Grid Search CV package. The grid search tuning was implemented for each ML across all provinces. The input data used for the grid search tuner was selected per province and a lag period of 5 days across all features were also used as input for each ML model.

The grid search tuning was implemented in the following steps:

1. Selection of a province.
2. Selection of an ML algorithm (CNNs, LSTMs, SVMs).
3. The climate features and diarrhoea cases data for a specific province and a lag of 5 days was selected as input.
4. Pre-processing for the chosen algorithm took place.
5. The selected parameters to be tuned for the chosen algorithm were tuned with the grid search tuner.
6. The best set of parameters were recorded.

See Table 4.2 (a, b, c) for the list of parameters we tuned for each ML model.

4.5.1.2. ML Model Configuration

- SVM Model - SVMs were one of the ML algorithms that was proposed for this study and are described in Chapter 3. For building the SVM daily diarrhoea cases prediction model, the python Scikit-Learn package for SVM was adopted in this study. An SVM with Radial Basis Function (RBF) kernel was used for prediction. An SVM with RBF kernel usually have two important parameters that influences performance. They are C and gamma (λ). To determine the optimal values for C and gamma (λ), grid search is used (see section 4.5.1.1 and Table 4.2(a)). The best parameters selected by the grid search tuner for a specific province was used to make predictions for that province across all lags. Before training, the data pre-processing technique explained in section 4.3 is implemented and after training all scales are reversed before evaluations are made on the model's prediction.

- LSTM Model – LSTMs were also used to make diarrhoea cases predictions and are described in Chapter 3. All LSTM models were implemented with the python Keras deep learning Library and TensorFlow backend. The models were configured to make reproducible results thus, a fixed random seed was set for all experiments. We used the Adam optimizer, mean square error loss function and Tanh activation function for all our experiments because studies have shown them to be very likely to achieve good results [16], [66] . Other parameters were selected with the grid search parameter tuner. See Table 4.2(b) for more details. The best parameters selected by the grid search tuner for a specific province was used to make predictions for that province across all lags. Before training, the data pre-processing technique explained in section 4.3 was implemented and after training all scales are reversed before evaluations are made on the model's prediction.

- CNN Model – CNNs were used to make diarrhoea cases predictions for each province and are also discussed in Chapter 3. All CNN models were implemented with the Keras deep learning Library and TensorFlow backend. The models were configured to make reproducible results thus, a fixed random seed was set for all experiments. We used the Adam optimizer, mean square error loss function and Relu activation function for all our experiments because they have been known to achieve accurate results for most CNN prediction studies [22]. Other parameters were selected with the grid search parameter tuner. See Table 4.2(c) for more details. The best parameters selected by the grid search tuner for a specific province was used to make predictions for that province across all lags. Before training, the data pre-processing technique explained in section 4.3 was implemented and after training all scales are reversed before evaluations are made on the model's prediction.

Table 4. 2: Grid Search and REVAC Parameter Boundaries for all SVM, LSTM & CNN Prediction Models

(a) Parameter Boundaries for the SVM Model

Parameter	***Parameter Range***
C	{1, 100}
λ	{0.001, 0.1}

(b) Parameter Boundaries for the LSTM Model

Parameter	***Parameter Range***
Neurons	{6,12,16,18,24,28,32,50,64,100}
No of epochs	{40,50,60,70,100,120,150,200}
Batch size	{4,16,18,32,64}
No of stacked LSTM layers	{1,2,3}
Learning rate	{0.001, 0.01}
Dropout rate	{0.1,0.2,1.0}
Optimizer	Adam (fixed)
Loss function	MSE (fixed)
Activation function	Tanh (fixed)

(c) Parameter Boundaries for the CNN Model

Parameter	***Parameter Range***
Convolution layers	{1,2,3}
Kernel size	[6,12,16,18,24,28,32,64]
No of epochs	{40,50,60,70,100,120,150,200}
Pool size	{1,2}
Batch size	{4,16,18,32,64}
Learning rate	{0.001, 0.01}
Optimizer	Adam (fixed)
Loss function	MSE (fixed)
Activation function	Relu (fixed)

4.5.1.3. Summary of Prediction Tasks Conducted for Experiment I

The set of tasks designed for each of the three proposed ML algorithms with respect to the original data (as explained in section 4.5.1) can be summarised in the following steps:

1. Selection of a specific province.
2. Selection of an ML algorithm (CNNs, LSTMs, SVMs).
3. The climate features and diarrhoea cases data for a specific province and a specific lag period (1 day, 5days, 2weeks, 3weeks) was selected as input.
4. Pre-processing for the chosen algorithm took place.
5. Each ML model was configured based on section 4.5.1.2 and the parameters for the algorithm were set based on the selection made by the grid tuner (see section 4.5.1.1)
6. Post-processing took place and predictions were made.

For each province, these steps were repeated 3 times for each lag for each ML algorithm across all province and the average RMSE result for each lag and each ML was stored. To determine the algorithm with the best performance, we compare their average RMSE across all lags per province.

4.5.2. Experiment II: Predictions with Original Data and Synthetic Data

The second set of experiments conducted for this study were implemented with a combination of the original data (see section 4.5.1) and synthetic data generated by GANs in different proportions across all lags. The aim of this experiment was to determine the effect of training data size (based on the combination of both synthetic and original data) to the prediction performance of all ML algorithms. This section tries to explain the experiments carried out to generate the synthetic data as well as the experiments for the ML predictions we made afterwards (that is, with both original and synthetic data).

4.5.2.1. Synthetic Data Generation

We used GANs to generate synthetic data for this study. This section explains how GANs were used to generate realistic diarrhoea and climate datasets. The synthetic data generated were used to augment the original datasets we had for training and testing.

- **Data pre-processing:** To train the GAN, the original daily diarrhoea and 8 climate features datasets with a sequence length of 24 was used per province. The datasets were normalized/scaled with the Min-Max scaler python Scikit-learn package with a feature range of (-1,1). After pre-processing, there was 3736 data instances across all 9 features available for training. After synthetic samples were generated, the datasets were reverted to their original scale.

- **GAN Architecture:** The GAN model we used made use of LSTM network for both the generator and the discriminator. The choice of the LSTM network was due to the fact that studies have shown them to be good for learning sequences [16], [19] and our training data was time series in nature. The LSTM network we used for our generator had a depth of 3 with 100 hidden units while our discriminator LSTM network had depth of 1 with 100 hidden units as well. Since GANs generate samples from a specific latent space, (a latent variable is an unobserved variable, and a latent space is a multi-dimensional vector space of these variables. The latent dimension is basically the size of the latent space) we tried different latent dimensions ranging from 5 to 70. We noted that larger latent space dimensions generate more realistic looking samples especially with multivariate datasets. The cross-entropy loss was used to measure the performance of the discriminator and generator.

- **Sample Generation:** To generate samples, the GAN model was trained with different batch sizes (8, 16, 32) across different epochs (200,300,400,500) and latent dimensions. To determine if the synthetic samples were close to the original data, visual comparison between the original and synthetic data was done. In addition, we computed the average difference measure between the original and synthetic data to further determine how close the synthetic data is to the original data. In the early stage of learning, the samples were different but as learning progresses further, the model eventually generates realistic looking samples for the diarrhoea and climate features dataset. After training, the GAN model was used to generate 20,000 synthetic samples. These samples were in the form of samples, timesteps (also known as sequence length) and features where each time step can be used as a lag period during prediction experiments. The GAN model was trained separately for each province and for each province, 20,000 synthetic samples with a sequence length of 24 was generated and used for our subsequent experiments.

4.5.2.2. Data Augmentation

The datasets used for training the GAN model were the daily diarrhoea cases and daily climate features datasets for all the 9 provinces. After training, twenty thousand synthetic data samples which had 24 timesteps for each of the 9 features (that is, diarrhoea cases and climate variables) were generated for each province. To prepare the dataset for predictions with the three ML models, a combination of the synthetic and original dataset was made for each province.

The data from both the original and synthetic set based on a specific lag period were augmented in the proportions shown in Table 4.3. The GAN model in this study does not generate Date as a variable rather it generates samples in the form of a series/sequence. Therefore, the two datasets (original and synthetic) was combined in two ways explained below.

1. **Upward Augmentation**: The original data was added to the top of the synthetic data. When the datasets are augmented this way, the training set will include a combination of the original and synthetic samples, but the test set will include only the synthetic datasets since it is at the bottom and it is also has a larger sample size.

2. **Downward Augmentation**: The original data was added to the bottom of the synthetic data. When the datasets are augmented this way, the training set will include mainly the synthetic datasets due to its quantity and the test set will include the original dataset since the original data was added to the bottom of the synthetic data.

For each distribution in Table 4.3, the datasets for the upward augmentation and downward augmentation were used separately by the three ML Models for prediction across each province. For example, if a dataset for a specific province is prepared with the 50/50 distribution in table 4.3, it means that 50% (10,000 samples) of the synthetic data and 50% (1881 samples) of the original data will be augmented in both upward and downward manner and will be used separately by an ML model both for prediction. In addition, after augmentation, 70 percent of the total datasets was used for training while the remaining 30 percent were used as test set. The datasets used for prediction per province per lag (that is, lag of 1, lag of 5, lag of 2weeks, lag of 3 weeks) in Experiment II are summarised in Table 4.3.

4.5.2.3. Summary of Prediction Tasks Performed in Experiment II

The set of prediction tasks conducted for Experiment II were implemented with a combination of the original data and synthetic data which both included diarrhoea cases and 8 climate features. Before training, the data for a specific province is pre-processed according to a lag period and a specific proportion in both upward and downward augmentation. For each province, all experiments were carried out across four different lags periods (see section 4.2) and across all proportions (see Table 4.3). This was done for each ML model for both upward and downward data augmentation. For all the prediction tasks carried out in Experiment II, the configuration for the three ML models were the same as the one used in Experiment I. The parameters selected by the grid search tuner for each ML algorithm and each province in Experiment I were also used for all the tasks in Experiment II. The procedure for the tasks carried out in Experiment II can be summarised in the following steps:

1. Select a specific province.
2. Select an ML algorithm (CNNs, LSTMs, SVMs) for prediction task.
3. Climate features and diarrhoea cases data for a specific province and a specific lag period (1day, 5days, 2weeks, 3weeks) and a specific proportion for a specific augmented data (that is, upward or downward augmentation) was selected as input.
4. Pre-processing for the chosen algorithm took place.
5. For each province, the selected ML model was configured based on section 4.5.1.2 and the parameters for the algorithm are set based on the selection made by the grid tuner in section 4.5.1.1.

For each province, these steps were repeated 3 times for each lag for each proportion and each combination for each ML algorithm and the average RMSE result was stored each time. To determine the algorithm with the best performance, we compare their average RMSE across all lags and proportions per province for upward and downward data augmentation separately.

Table 4. 3: Distribution and proportions of dataset used for prediction for each features and provinces

Synthetic Samples (size =20,000)	*Original Samples (size = 3,763)*
90% (18000)	10% (376)
80% (16,000)	20% (753)
70% (14,000)	30% (1130)
60% (12,000)	40% (1505)
50% (10,000)	50% (1881)

4.5.3. Experiment III: Predictions with Original Data and Synthetic Data and REVAC Parameter Tuning

The third set of experiments conducted for this study were implemented with a combination of the original data and synthetic data generated by GANs in different proportions (see section 4.5.2.2). The aim of this experiment was to determine the effect of REVAC parameter tuning on the performance of the predictions made with a combination of both the original and synthetic dataset. The major difference between Experiment II and Experiment III is the method used for tuning the parameters of each ML model. In Experiment III, the parameters of each ML model were tuned with REVAC tuning algorithm explained in the Method section (see section 3.4). This section tries to explain the how REVAC tuning was used to tune the parameters of all the ML algorithms we used for the daily diarrhoea case predictions.

4.5.3.1. REVAC Tuning

The REVAC algorithm adopted for this study was based on the methodology used by Nannen & Eiben [34]. REVAC was implemented at a layer that aids in searching for optimal parameter values for an ML algorithm trying to solve the problem of predicting daily diarrhoea cases. The input data used for this task were divided into three parts. In other words, they were used separately for every REVAC tuning iteration per province. They are:

- The original data
- The combination of original and synthetic in the 50/50 proportion for upward dataset augmentation (see section 4.5.2.2 and Table 4.3).
- The combination of original and synthetic in the 50/50 proportion for downward dataset augmentation (see section 4.5.2.2 and Table 4.3).

For each of the above input data, a lag period of five days across all features was used as input data for tuning each ML algorithm. The objective of the REVAC tuner was to minimize a given fitness function. The fitness function of this experiment was calculated as the RMSE of the predictions made by the ML algorithm for each of the above input data. The best parameters yielded by each ML algorithm for each of the above input data based on the REVAC tuner was stored and used later for final predictions.

REVAC itself works with a set of parameters that determines how efficiently it runs. The list of REVAC parameters and their values can be seen in Table 4.4. These values were chosen based on the recommendations of the paper by Nannen & Eiben [34]. To tune the parameters of each ML algorithm these REVAC parameters must be set first. The list of parameters for each ML algorithm to be tuned were the same as the parameters we tuned with the grid search tuner in Experiment I. Table 4.2 contains this list of ML parameters we tuned with REVAC. The REVAC tuning task can be summarized in the following steps:

1. A generation size of hundred and an initial population size of 80 was set.
2. Select a province.
3. Select an ML algorithm (CNNs, LSTMs, SVMs) to be tuned.
4. One of the 3input data (explained above) for a specific province and a lag of 5 days was selected as input.
5. Pre-processing for the chosen algorithm took place.
6. The chosen parameters to be tuned for the selected algorithm were tuned with the REVAC tuner.
7. After hundred generations, the set of parameters with the best fitness (that is, least RMSE) is recorded.
8. Repeat these steps separately for the other two sets of input data explained above.

This means that for each province, the REVAC tuner will yield a set of fittest parameters for the three set of input data for each ML algorithm. Figures 4.1 to 4.3 show a heatmap of how the RMSE error changes after each generation during REVAC tuning for each algorithm. These figures are for the North West province's original data only. By observing these figures, we notice that in the early generation, large RMSE are yielded but as time progresses, the RMSE becomes smaller and the algorithm converges after at least 50-60 generations for the deep learning models. The SVM model on the other hand converges earlier than 50-60 generations.

4.5.3.2. Summary of Prediction Tasks Performed in Experiment III after REVAC Tuning

After the REVAC parameter tuning tasks have been completed, three fittest set of parameters were recorded for the three sets of input data was used for the REVAC tuning tasks in section 4.5.3.1. These parameters were used to carry out final predictions and

the best results were selected. The datasets used for predictions in Experiment II were the same used for predictions in Experiment III.

For all the prediction tasks carried out in Experiment III, data pre-processing steps and the configuration for the three ML models were the same as the one used in Experiment I. However, the parameters we used for these experiments were the parameters selected by the REVAC tuner. Since the REVAC tuner selected three sets of fittest parameters for each province based on three set of input data, we had to choose one of these set of parameters for a specific province for our final prediction. To achieve that, the following steps were taken:

1. Select a specific province.
2. Select an ML algorithm (CNNs, LSTMs, SVMs) to make prediction.
3. Climate features and diarrhoea cases data for a specific province and a specific lag period (1day, 5days, 2weeks, 3weeks) and a specific proportion for a specific augmented data (that is, upward or downward augmentation) was selected as input.
4. Pre-processing for the chosen algorithm took place.
5. If upward data augmentation dataset was selected as input, the ML model parameters was configured based on the first set of fittest REVAC parameters (that is, either the original or upward dataset parameters selected by REVAC tuner was used. Downward dataset fittest parameters were used instead of upward when downward data augmentation was selected as input).
6. The predictions were made thrice and the average RMSE was stored.
7. Steps 1-6 were repeated for each lag, each proportion, and each dataset augmentation for each ML algorithm over each province.
8. For each ML algorithm and each province, the average RMSE across all lags and all proportions for each dataset augmentation were calculated and stored.
9. Steps 1-8 above were repeated with the second set of fittest parameters depending on which set of fittest parameters were used in step 5.
10. The two final RMSE average based on the first set of fittest parameters and second set of fittest parameters will be compared and the one with the least RMSE was recorded while the other was discarded.
11. To determine the algorithm with the best performance, we compare their average RMSE across all lags and proportions per province for upward and downward data augmentation separately.

4.6. Statistical Analysis Performed for all Experiments

We conducted some statistical tests to make robust conclusions about our research objectives. We drew up some hypothesis based on our research objectives in section 1.2 and experiments in Table 4.1. The hypothesis we drew up were:

1. Model performance are similar across each province and over all provinces.
2. The use of REVAC parameter tuning during training is similar to the Grid search parameters.

The first hypothesis stated in 1 above addresses the main objective of our study which is "to detect which supervised machine learning techniques (CNN, LSTM and SVM) performs best in terms of high accuracy when predicting number of diarrhoea cases given a range of datasets (for example, varying proportions of real and synthetic climate variables and diarrhoea datasets) for training and testing". It addresses this objective by checking if the performance of the ML models is similar. It achieves this by testing statistical significance between the prediction results of each ML model against the other (that is, CNN vs LSTM, LSTM vs SVM and CNN vs SVM) for each province and over all provinces (average results across all province.) in Experiment I, II and III respectively. The data used for conducting these tests can be seen in Table 5.1-5.2 (results for Experiment I) Table B1-B18 in Appendix B (results for Experiment II) and Table C1-C18 in Appendix C (results for Experiment III).

The second hypothesis stated in 2 above addresses the objective which states that "investigate to what extent REVAC parameter tuning can improve the accuracy of the three models". It addresses this objective by checking if the results of a ML model are similar when either of the tuning methods are used. It achieves this by testing statistical significance between the results of a specific ML model when grid search tuning was used for predictions against its results when REVAC tuning was used for predictions in each province and over all province (that is performance of a specific ML model in Experiment II vs its performance in Experiment III). The data used for conducting these tests can be seen in Table B1-B18 in Appendix B (results for Experiment II) and Table C1-C18 in Appendix C (results for Experiment III).

Table 4. 4: Parameter used for REVAC Tuning and all SVM, LSTM & CNN Prediction Models

Parameter	***Parameter Range***
Population Size	80
No. of Generations	100
No. of Parents for Crossover	2
No. of children to be created per generation	1
No. of parents to be replaced per generation	1

For each hypothesis, the statistical tests we conducted were the Shapiro-Wilke test [89] to test if the outcomes of each experiment was normally distributed and the Wilcoxon signed ranked test [90] to test for significance between our results in Experiment I, II and III. Shapiro-Wilke test was used because of the small sample size of the datasets to be tested. We chose the Wilcoxon signed ranked test because it is appropriate for non-normally distributed samples that are related [91]. We test for significance by running these tests between the final average RMSE results for ML Model against the other (between each province and over all province) in Experiment I, II and III separately, and results between the parameter tuning methods used during training. See Appendix A for the outcomes of our statistical tests.

For all tests, we regarded $p < 0.05$ as being statistically significant because it is considered as the reasonable standard of significance in most scientific research [92]. The statistical tests were applied in a pairwise manner and the Benjamin Hochberg [93] correction test was applied to decrease the false discovery rate for multiple comparisons since multiple statistical tests were conducted between the average results of each ML model in each province and over all province. All the statistical and correction tests were conducted with the inbuilt statistical library for R programming language.

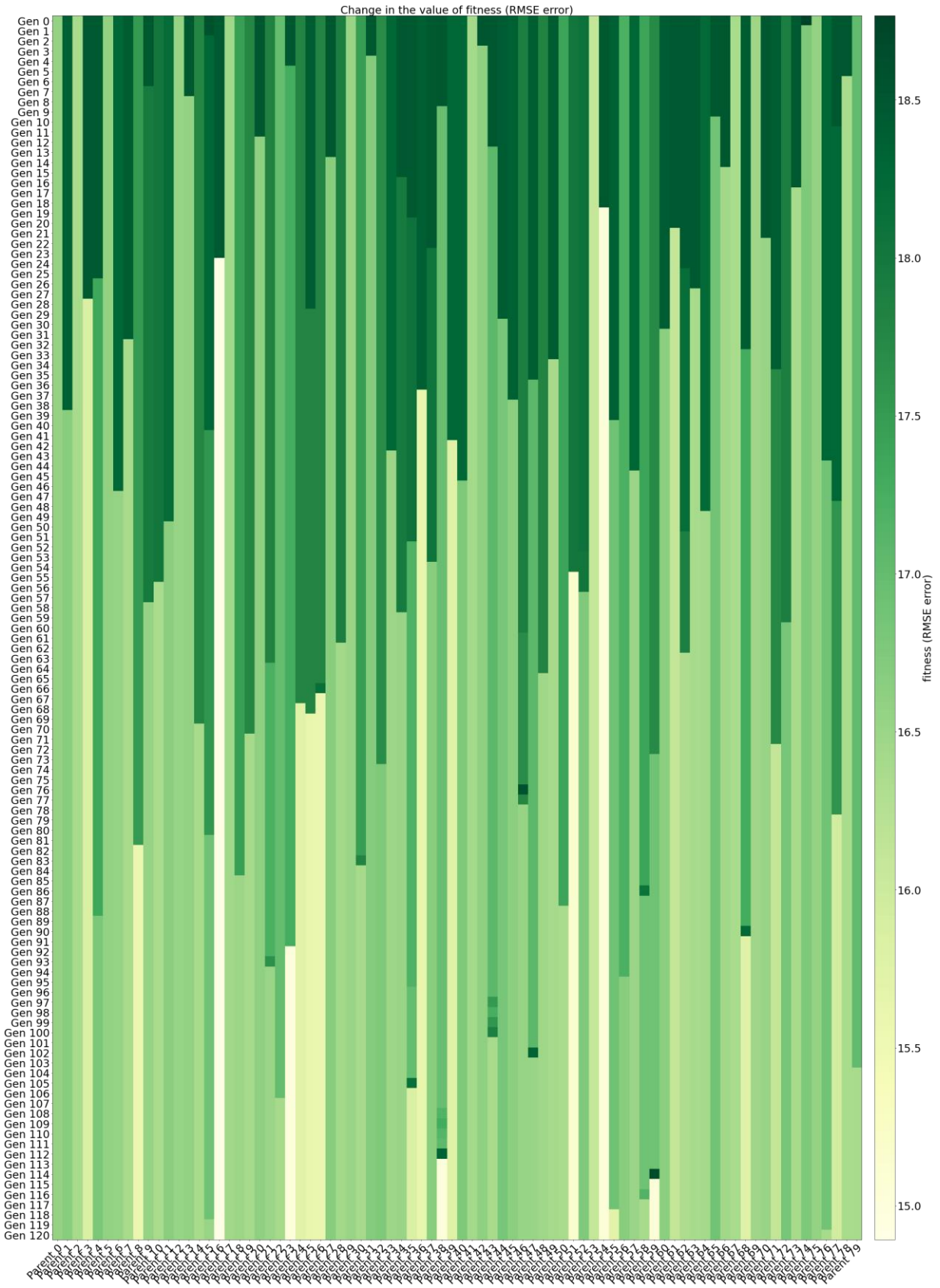

Figure 4. 1: *Heat Map showing changes in RMSE scores using REVAC for the SVM model. The original dataset for North West Province was used as input for this REVAC run. y-axis represents number of generations and x-axis represents parents in the population.*

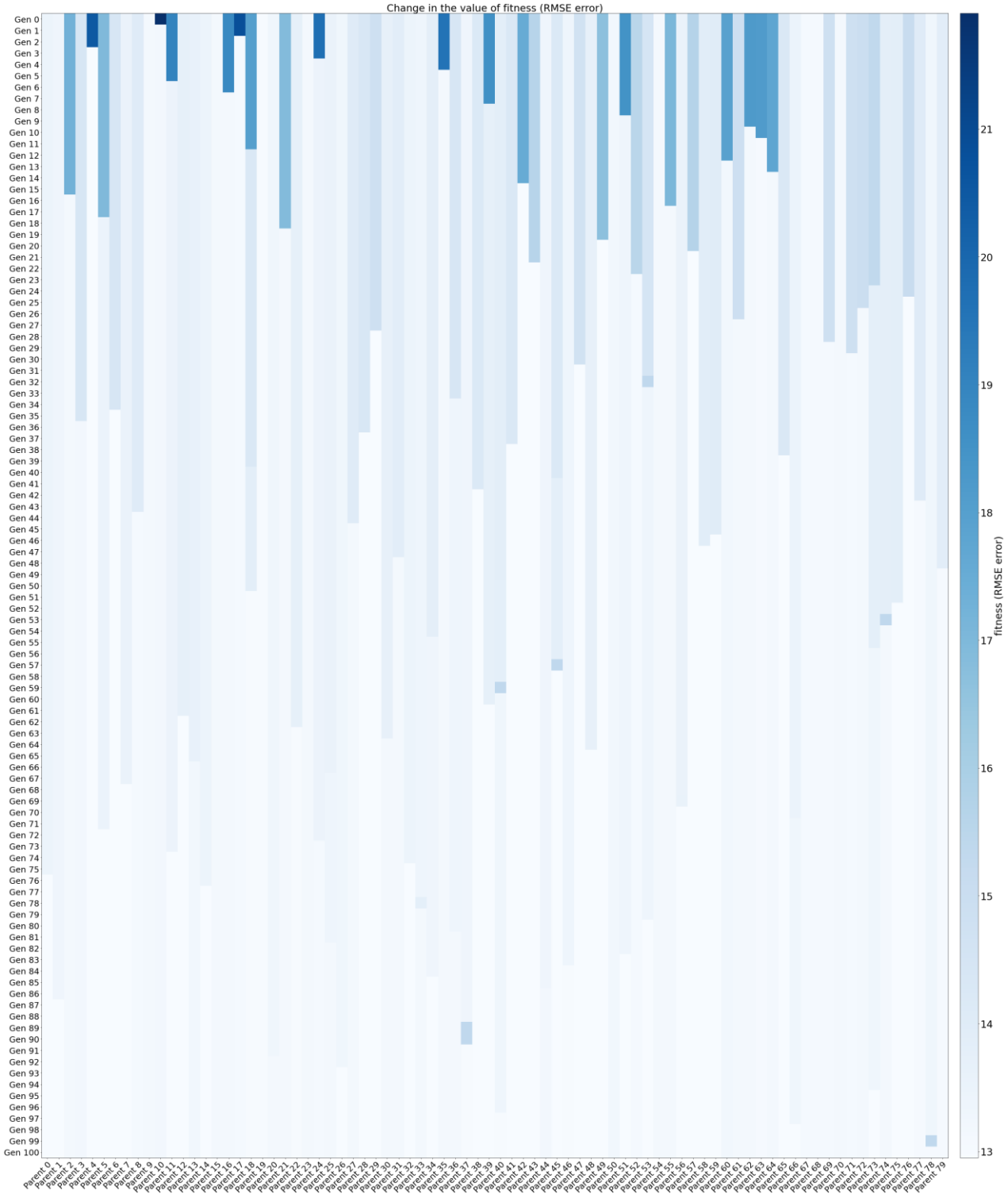

Figure 4. 2: *Heat Map showing changes in RMSE scores using REVAC for the CNN model. The original dataset for North West Province was used as input for this REVAC run. y-axis represents number of generations and x-axis represents parents in the population.*

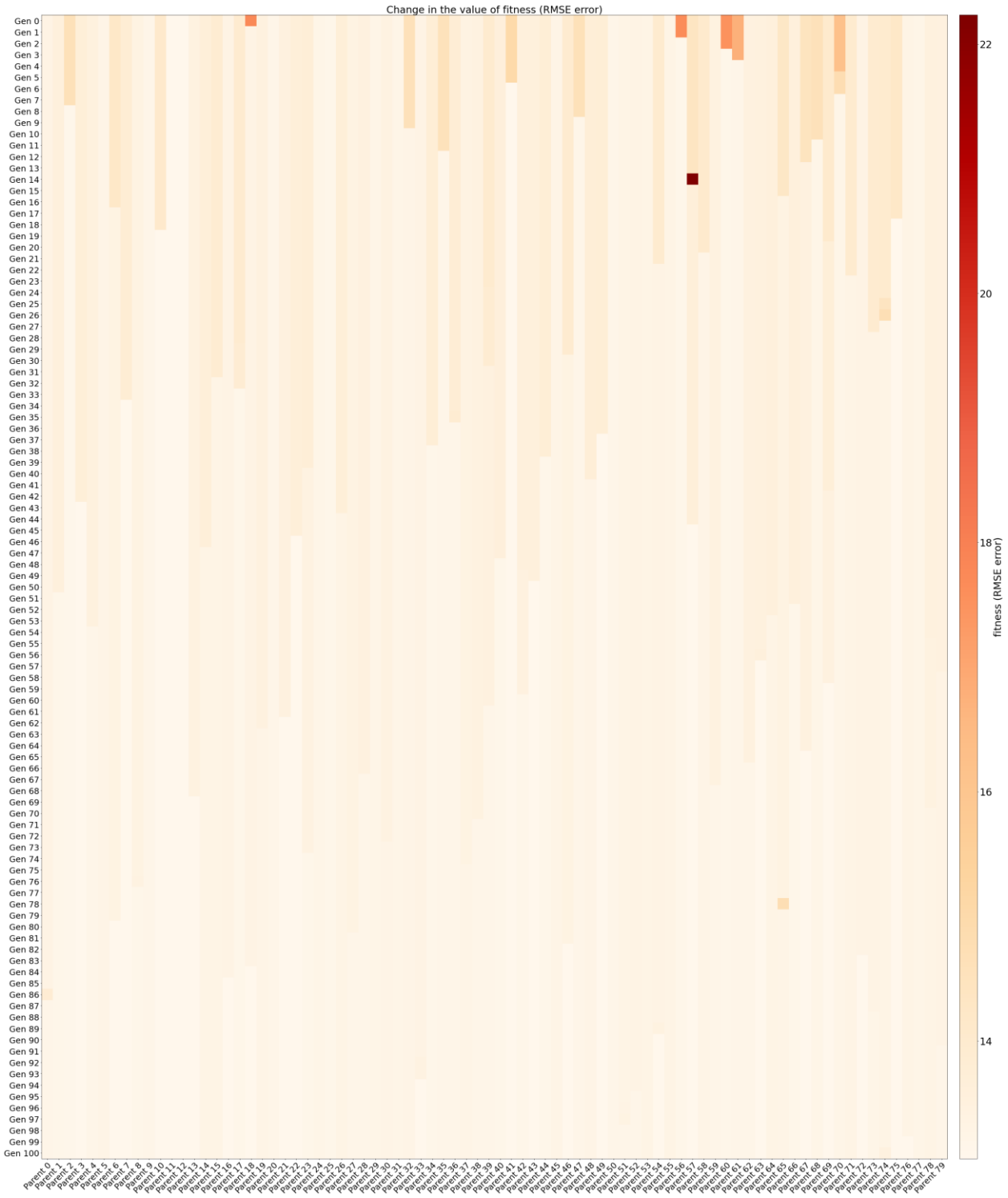

Figure 4. 3: *Heat Map showing changes in RMSE scores using REVAC for the LSTM model. The original dataset for North West Province was used as input for this REVAC run. y-axis represents number of generations and x-axis represents parents in the population.*

4.7. Sensitivity Analysis

To measure the degree of importance of each climate variable to the diarrhoea prediction model in a specific province, we conducted a sensitivity analysis [8] to examine the contribution of each climate variable to the output of the best predicting model for each province in Experiment I. The CNN model was used to conduct the sensitivity test since it outperformed the other models in almost all provinces. We adopted the Backward stepwise method [94] in which we measured the effect of one variable at a time while keeping the other variables fixed. Sensitivity is then measured by observing changes in the RMSE error of the model based on the omission of a certain variable. The larger the increase in RMSE, the higher the importance of the omitted variable.

4.8. Summary

This chapter discussed how the experiments for this study was conducted in order to achieve our research objectives. It gave details on the datasets we used, the performance evaluation criteria and the experiments performed to determine the best performing algorithm. The section also described how our ML algorithms were configured including the parameter tuning strategies we used to obtain optimal parameter values. Finally, the sections explained how statistical tests were conducted to make robust conclusions as well as the sensitivity study that was used to determine level of importance of each climate variable to the prediction model in each province.

Chapter 5

5. Results

This section provides the results of our experiment sets in Table 4.1. It also describes the outcomes of the research objectives outlined in section 1.2. In this study, we used quantitative measures to assess the performance of our predictive models. Although graphical representations provided us with some inference for our experiments' results, we went further to conduct some statistical test to make robust conclusions. We used the Wilcoxon signed ranked statistical test to test for significance (see section 4.6 for details). Our findings are presented in the following sections.

5.1. Outcomes for Experiment I

This section shows the results for the predictions tasks carried out with the original data in Experiment I (see Table 4.1 for details). The results in this section aim to address the first objectives in section 1.2 which state that: *"Test the performance of existing deep learning methods such as CNNs, LSTMs and an existing conventional ML method like the SVMs across a range of datasets (that is, varying proportions of real and synthetic climate variables and diarrhoea-based datasets at different testing and training intervals)."*

To address the objective above, we compared the RMSE of the three ML methods based on the predictions made in Experiment I. Figure 5.1, represents the average RMSE results for the prediction tasks conducted in Experiment I for each province. In this figure, we observed that CNN's predictions outperformed the other models in almost every province with the exception of Limpopo province where the LSTM model outperformed the others. Over each province, the performance of the CNN model was closely followed by the LSTM model while the SVM, with the highest RMSE error, showed the poorest performance. However, in Limpopo province the LSTM RMSE average was 9.95 while CNN and SVM averages were 10.27 and 11.00, respectively. Over Western Cape province, CNN had the least RMSE average of 85.42 while the LSTM and SVM averages were 90.47 and 90.58, respectively. We also observed that the difference margin in RMSE between the three models was larger in both Western Cape and Gauteng province and least in Mpumalanga, Free State, and Northern Cape province.

Figure 5.2 compares the RMSE averages of the three ML models over all provinces. We observed that CNN had the least overall RMSE average of 31.55 while LSTM and SVM averages were 32.91 and 33.89, respectively. We can infer from these results that the RMSE errors are lower for deep learning models (CNN & LSTM).

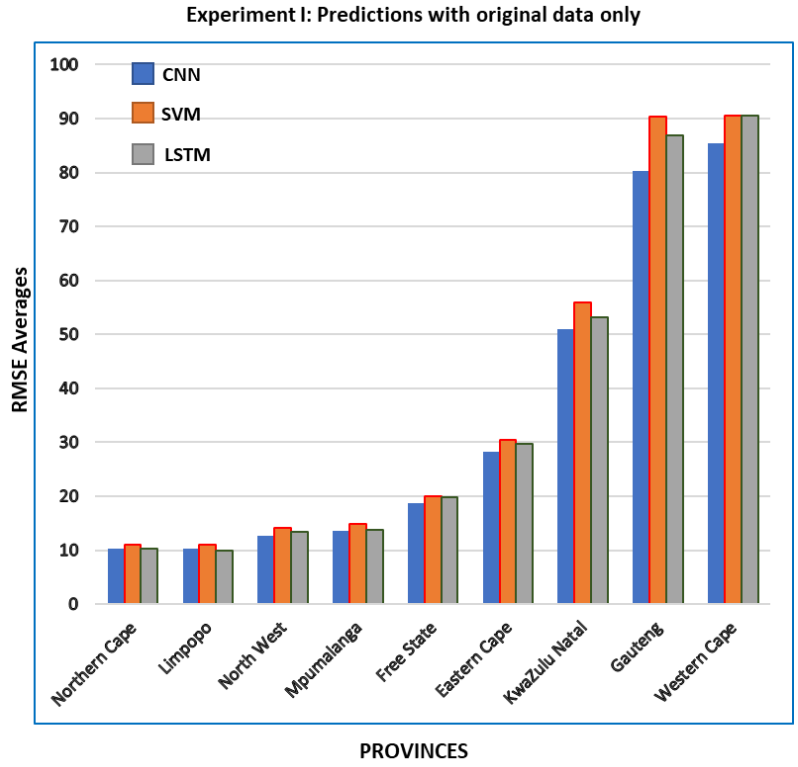

Figure 5. 1: *CNN, SVM, LSTM average RMSE errors for all prediction scenarios in each province for Experiment I (see Table 4.1 for details on Experiment I). Lower RMSE averages indicate better performance.*

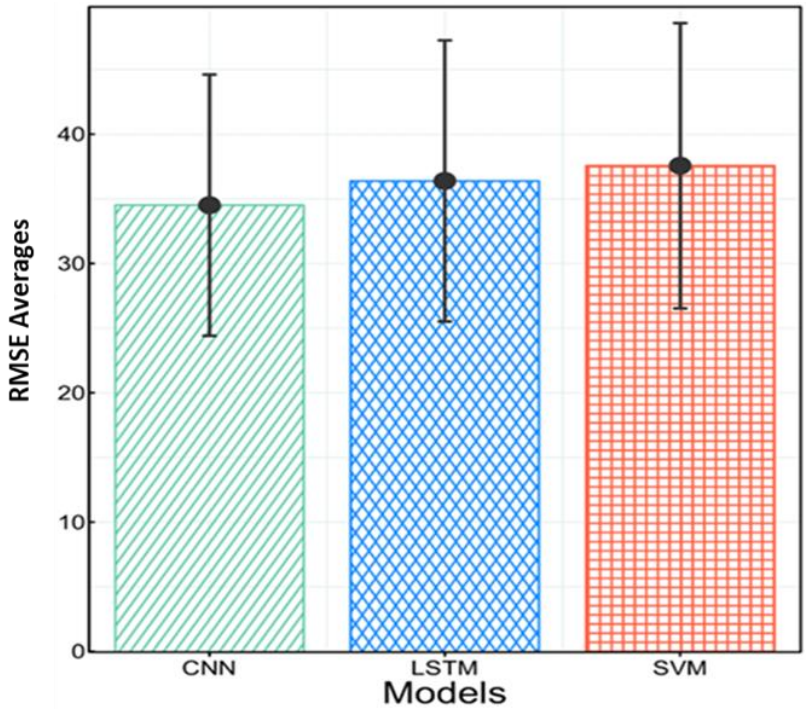

Figure 5. 2: *CNN, LSTM, SVM average RMSE errors over all provinces for all prediction scenarios in Experiment I. Low RMSE average indicates better performance accuracy. (See Table 4.1 more details on Experiment I). The arrows represent the corresponding widths of twice the standard error.*

Although graphical representations in figures 5.1 and 5.2 provided us with some inference for the experiments, statistical test was performed to make robust conclusions. Therefore, to further address the objective stated above, we formulated hypothesis zero (H0) which states that*: "Model performance are similar within the original dataset across each province and over all provinces".* Thus, we applied multiple Wilcoxon signed rank test in pairwise comparisons with the Benjamin Hochberg correction test (see section 4.6 for details) between the average RMSE results for the three ML models. Table A.1 in the Appendix shows the outcomes of the statistical test conducted for all possible comparisons in Experiment I. We observed that there was no statistically significant difference between all the model comparison for each of the nine provinces. However, when we compared the average RMSE across the nine provinces (altogether), the performance margin was statistically significant for comparisons between the SVM and CNN models and comparisons between LSTM and CNN models. The lack of significant difference between the models when compared per province may be due to the small sample size used during the test [92]. For all prediction scenarios in Experiment I, we used different lag variables to see how long-term trends of the climate features affect the prediction performance of our models.

Table 5. 1: RMSE errors from the CNN, SVM and LSTM model for all prediction scenarios with the original dataset in Experiment I. Lower RMSE indicate better prediction accuracy of the model and vice-versa.

Provinces	*CNN RMSE Original (Lag 1)*	*SVM RMSE Original (Lag 1)*	*LSTM RMSE Original (Lag 1)*	*CNN RMSE Original (Lag 5)*	*SVM RMSE Original (Lag 5)*	*LSTM RMSE Original (Lag 5)*
Western Cape	106.35	115.67	107.03	83.16	86.13	83.93
KwaZulu Natal	59.71	62.85	64.05	49.91	51.79	46.38
Limpopo	11.44	13.03	11.27	9.58	10.29	9.40
Free State	21.47	24.82	21.32	18.24	19.77	18.55
Mpumalanga	14.45	17.02	14.53	13.56	14.33	13.60
Northern Cape	10.15	11.70	10.24	10.30	11.07	10.17
North West	13.43	15.81	14.41	12.95	14.89	13.05
Gauteng	94.82	99.58	90.62	80.28	84.96	82.74
Eastern Cape	32.23	35.83	33.09	28.27	29.30	28.84

Table 5. 2: RMSE errors from the CNN, SVM and LSTM model for all prediction scenarios with the original dataset in Experiment I. Lower RMSE indicate better prediction accuracy of the model and vice-versa.

Provinces	*CNN RMSE Original (Lag 14)*	*SVM RMSE Original (Lag14)*	*LSTM RMSE Original (Lag14)*	*CNN RMSE Original (Lag 21)*	*SVM RMSE Original (Lag21)*	*LSTM RMSE Original (Lag21)*
Western Cape	76.99	80.00	85.28	75.17	80.51	85.64
KwaZulu Natal	46.81	53.43	53.56	47.67	55.64	48.57
Limpopo	9.95	10.06	9.80	10.12	10.63	9.31
Free State	17.39	17.73	19.55	17.45	17.42	19.77
Mpumalanga	13.37	14.07	13.59	13.31	14.26	13.57
Northern Cape	10.13	10.54	10.32	10.36	10.56	10.32
North West	12.16	12.83	12.74	12.35	12.79	13.15
Gauteng	72.50	85.25	86.77	73.82	91.32	87.60
Eastern Cape	26.59	28.21	28.45	25.92	28.17	28.68

In Table 5.1 & 5.2, we observed that for all provinces and models, lagging variables by only 1 day does not yield the best performance. In Western Cape, lagging the variables by 5 or 21 days yielded the best performance for the three ML models while in Eastern Cape and Mpumalanga, lagging the variables by 14 or 21 days also yielded better performance. In KwaZulu Natal and Limpopo province, lagging the variables by 5 days yielded the best performance while in Free State and Northwest, lagging variables by 14 days yielded better performance. In Gauteng province, lagging the variables by 5 or 21 days gave the best performance. However, in Northern Cape, lagging the variables made almost no different in model performance. This shows that precursory effect of the climate variables affects diarrhoea cases for each province in different ways

5.2. Outcomes for Experiment II

This section shows the results for the predictions tasks conducted with combinations from the original data and synthetic data in Experiment II (see table 4.1 for details). The results in this section aim to address the first objectives above as well as the second objective in section 1.2 which was to: *"Investigate the effect of the varying proportions of real and synthetic training and testing data on model performance in terms of prediction accuracy of the three models."*

To address the objective stated above, we compared the average RMSE of each model based on the prediction scenarios in Experiment I with the average RMSE for the prediction scenarios in Experiment II per province. Recall that the dataset used in Experiment I was only the original data while the dataset used in Experiment II were combinations of the original and synthetic data (see Table 4.1 for details). In addition, the three ML algorithms in both Experiment I and II used Grid search parameters (see Table 4.1 for details).

Figure 5.3 represents the percentage change in performance of each ML model when the combinations of the synthetic and original dataset (augmented upwards) were used in place of the original dataset in Experiment I. We found that the use of the upward combinations of the original and synthetic datasets greatly improved the performance of all the three ML models. Predictions for Limpopo province recorded the highest improvement with over 60% increase for each of the three models while Northern Cape province predictions recorded the least percentage increase with approximately 15%, 25% and 22% for CNN, SVM and LSTM models, respectively. Over all provinces, the percentage increase in performance for predictions of the LSTM and SVM model was more than the CNN model.

Figure 5.4 represents the percentage change in performance of each ML model when the combinations of the synthetic and original dataset (augmented downwards) were used in place of the original dataset in Experiment I. Similar to the results of the upward augmentation in figure 5.3, the performance of all three models also increased considerably with Limpopo's prediction recording the highest performance increase. However, the percentage increase in performance for Limpopo's predictions was approximately 50% for each of the three ML models. Over most provinces, the percentage increase in the predictions for the LSTM and SVM models was more than the CNN model. However, CNN's percentage increase in performance was more than SVM's in Western Cape and KwaZulu Natal provinces. We can infer from these results that the amount of training data used for training, significantly affects the prediction performance of all the three ML models.

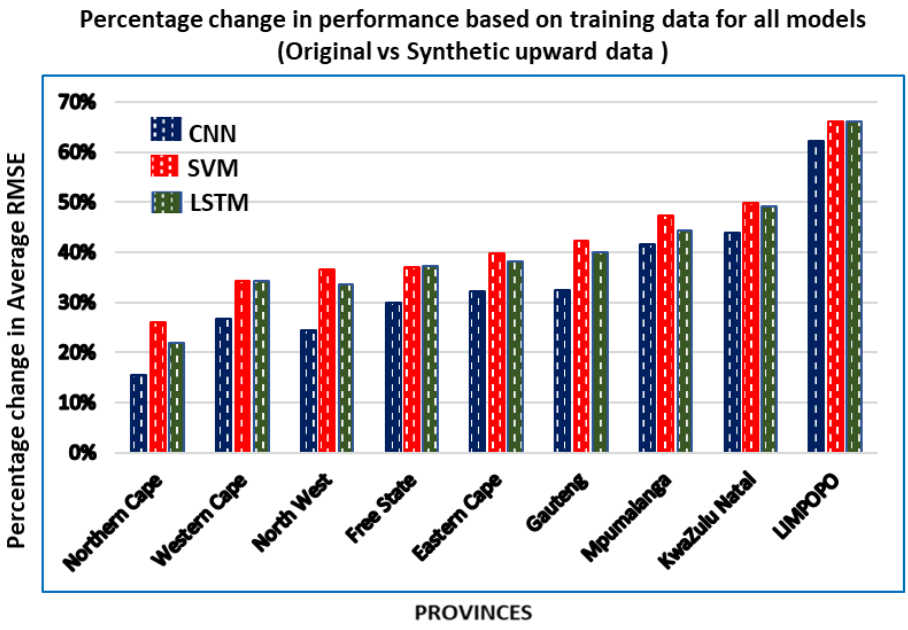

Figure 5. 3: *Percentage change in performance with (combinations of synthetic & original data augmented upwards) and without synthetic (original data only) training data for all three ML algorithms (CNN, SVM & LSTM) prediction scenarios conducted in Experiment II (see Table 4.1 for details on Experiment II). High percentages indicate high improvement in performance.*

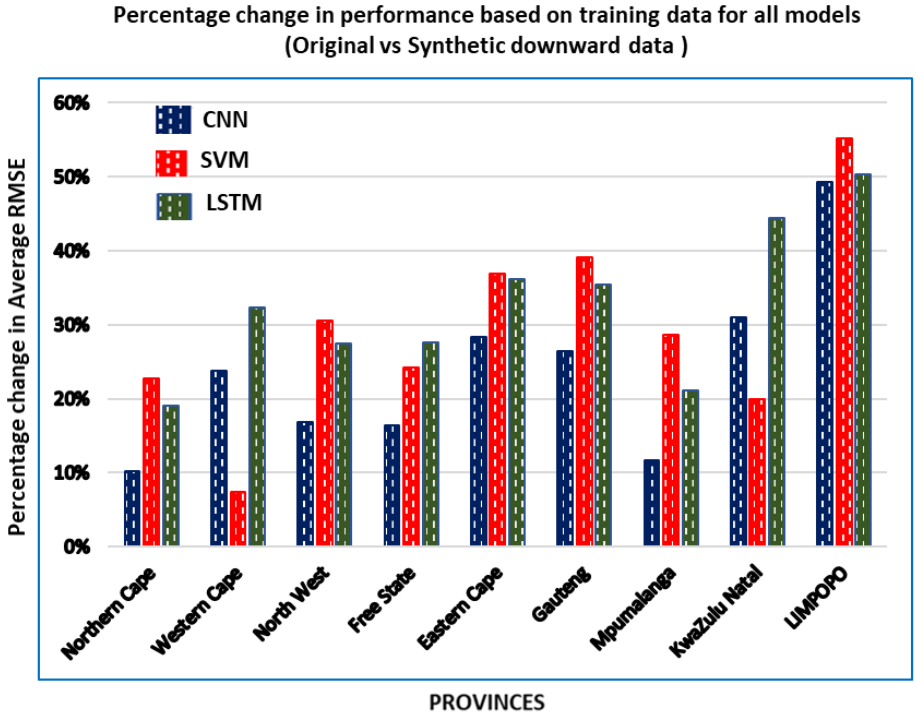

Figure 5. 4: *Percentage change in performance with (combinations of synthetic & original data augmented downward) and without synthetic (original data only) training data for all three ML algorithms (CNN, SVM & LSTM) prediction scenarios conducted in Experiment II (see Table 4.1 for details on Experiment II). High percentages indicate high improvement in performance.*

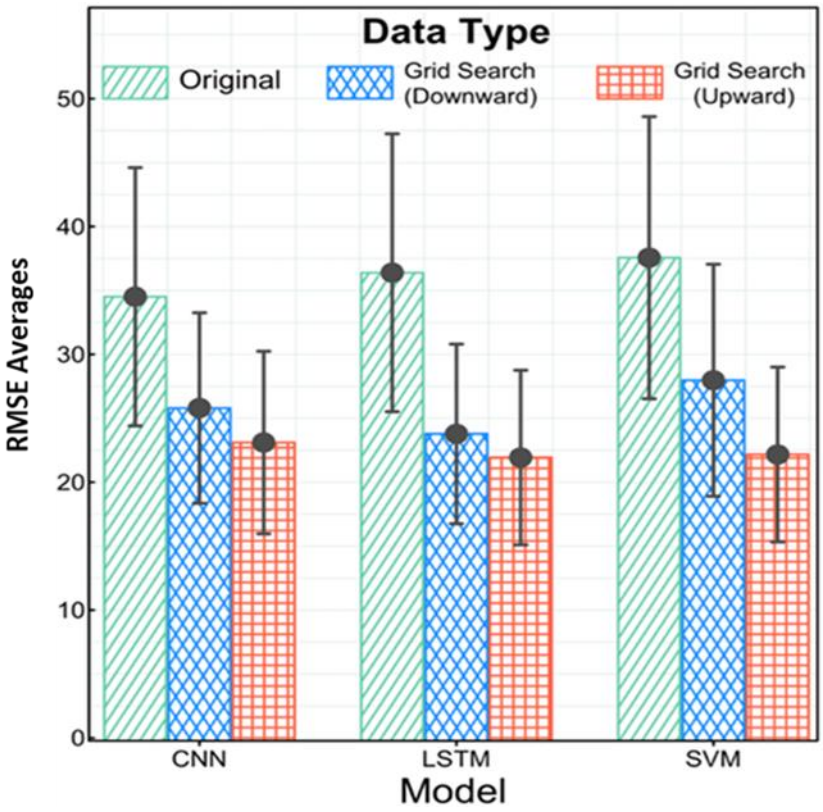

Figure 5. 5: *A comparison of CNN, LSTM, SVM average RMSE over South Africa (all provinces) for all prediction scenarios with the original data in Experiment I and all prediction scenarios with the downward augmented data and upward augmented data in Experiment II. Recall that Grid search was used in tuning the parameters of all ML models in both Experiment I & II. Low RMSE average indicates better performance accuracy. The arrows represent the corresponding widths of twice the standard error. See Table 4.1 for more details on both experiments.*

Figure 5.5 shows a further attempt to compare the performance of the overall predictions (that is, the prediction results over all provinces) made by each model using the original dataset in Experiment I, the upward augmented datasets in Experiment II and the downward augmented dataset in Experiment II. For each model, we also found that the predictions made with the upward augmented datasets yielded lower RMSE than their predictions with downward augmented datasets. For the CNN model, the overall RMSE for predictions made with the original, upward augmented, and downward augmented datasets were 31.55, 23.11 and 25.80, respectively. For each dataset, the SVM model's overall RMSE were 33.89, 22.17 and 27.97, respectively while the LSTM model's results were 32.91, 21.93 and 23.78, respectively. These results show that CNN outperformed the other models when the original dataset was used alone while LSTM outperformed the other models when synthetic datasets were augmented with the original dataset either upwards or downwards. We can also infer that all models benefitted from the increase in the quantity of datasets used for making prediction.

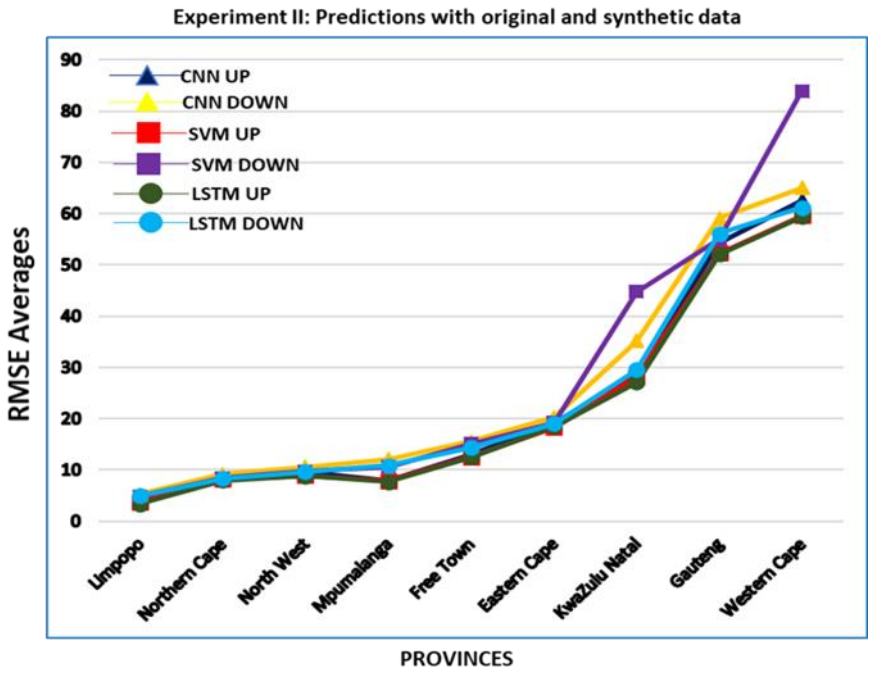

Figure 5. 6: *CNN, SVM, LSTM average RMSE error for all prediction scenarios in each province for Experiment II (see Table 4.1 for details on Experiment II). Lower RMSE averages indicate better performance.*

To further address the first objective in section 1.2, Figure 5.6 compares the average RMSE of the three ML models for both upward and downward augmented dataset predictions conducted in Experiment II for each province.

By comparing the models based on the predictions for the upward dataset combinations, over all provinces, LSTM model outperformed all the other models except in Eastern Cape province where SVM model yielded lower RMSE averages. For the other provinces, the results of the SVM model closely followed LSTM's however, CNN model outperformed the SVM model for predictions in Western Cape. When the models were compared based on their predictions for the downward dataset combinations, LSTM's prediction results still outperformed the other models for most provinces except in Limpopo, Mpumalanga, and Gauteng provinces where SVM outperformed both the LSTM and CNN models. However, both LSTM and CNN model's prediction results outperformed the SVM model in Western cape and KwaZulu Natal provinces by a very wide margin as illustrated in Figure 5.6. Although the charts in Figure 5.6 provided us with some useful inference, we went further to conduct statistical test to make robust conclusions as to whether the differences in prediction performance between models were statistically significant.

Due to the fact that we were comparing the performance between models across each province and over all province, we used hypothesis zero (H0) as our null hypothesis (see section 4.6 for details). We applied multiple pairwise Wilcoxon signed rank test with Benjamin Hochberg correction test (see section 4.6 for details). Table A.2 in the Appendix show the outcome of the statistical test conducted for all possible comparisons between the three model's prediction result in Experiment II for each province and over all provinces when averaged. The p-values that are bolded and asterisked represents statistically significant difference in performance between models. The lack of significance between models may be due to the small magnitude of the difference in RMSE. For further details of the analysis carried out in Experiment II see results in Tables B.1-B.18 in Appendix B.

5.3. Outcomes for Experiment III

This section shows the results for the predictions tasks conducted with combinations of the original data and synthetic data in Experiment III. The difference between Experiment II and Experiment III was the parameter tuning method the ML algorithms. In Experiment III, REVAC parameter tuning was used while in Experiment II, Grid search tuning was used (see Table 4.1 for details). The results in this section aims to address the first objective in section 1.2 as well as the third objective which is to: *"Investigate to what extent REVAC parameter tuning can improve the accuracy of the three models."*

To address the objectives stated above, we compared the average RMSE of each model based on the prediction scenarios in Experiment II with the average RMSE for the prediction scenarios in Experiment III per province. Figure 5.7 represents the percentage change in performance of each ML model when predictions for the upward augmented dataset are made with the parameters from REVAC tuning instead of the Grid search parameters. Our analysis show that the CNN model's prediction result improved across all the provinces especially in Northern Cape province where a percentage increase of over 8% was recorded. The prediction accuracy of the SVM model slightly improved for some provinces with Mpumalanga province recording the highest increase of about 1.3% however, prediction results for provinces like Western Cape and Free State provinces had negative percentages of over 1%. The prediction accuracy for the LSTM models also increased over most provinces with the highest increase of over 3.5% recorded for Free State's predictions. In Limpopo province, the LSTM's prediction performance declined drastically by over 6%. Prediction accuracy for KwaZulu Natal province also declined by over 1%.

Figure 5.8 represents the percentage change in performance of each ML model when predictions for the downward augmented dataset are made with the parameters from

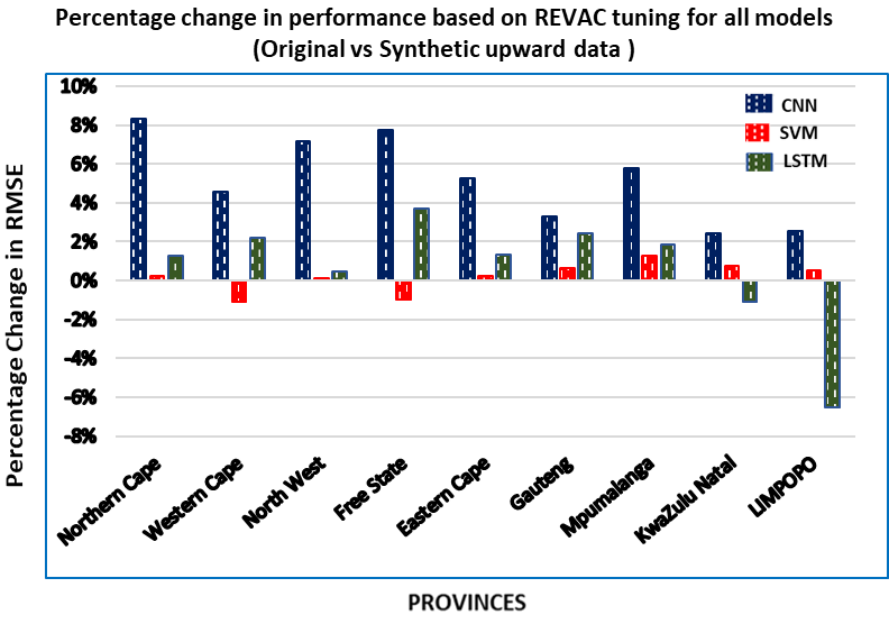

Figure 5. 7: *Percentage change in performance for all three ML algorithms (CNN, SVM & LSTM) prediction scenarios in Experiment III with (REVAC parameter tuning during training) compared with the results in Experiments II (without REVAC tuning). (Data used for training scenarios were combinations of synthetic & original data augmented upwards). High percentages indicate high improvement in performance. See Table 4.1 for more details on both experiments.*

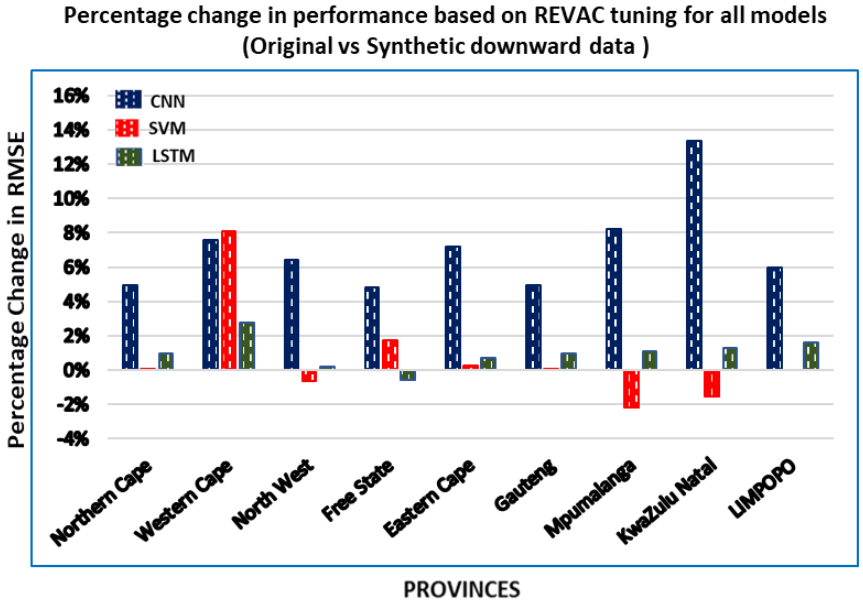

Figure 5. 8: *Percentage change in performance for all three ML algorithms (CNN, SVM & LSTM) prediction scenarios in Experiment III with (REVAC parameter tuning during training) compared with the results in Experiments II (without REVAC tuning). (Data used for training scenarios were combinations of synthetic & original data augmented downwards). High percentages indicate high improvement in performance. See Table 4.1 for more details on both experiments.*

REVAC tuning instead of the Grid search parameters. We found that the CNN model's prediction result improved once again across all the provinces especially in KwaZulu Natal province where a percentage increase of over 13% was recorded. The least percentage increase recorded for CNN model was about 5%. SVM's prediction accuracy on the other hand increased drastically for the Western Cape province with about 8%. Provinces like Limpopo, Gauteng, Eastern Cape and Northern improved very slightly while North West, Mpumalanga and KwaZulu Natal provinces reduced in prediction accuracy with Mpumalanga's result recording the highest decline of about 2%. The LSTM's prediction accuracy improved slightly over all provinces with the exception of Free State's province prediction where a decrease of about 0.6% was recorded. The highest increase of about 3% was recorded for Western Cape province. From figures 5.7 and 5.8, we can infer that the CNN model's predictions benefitted the most form REVAC parameter tuning when compared to the LSTM and SVM models.

Although Figures 5.7 & 5.8 provided us with some inference on the percentage change in performance when REVAC was used for parameter tuning, to make robust conclusions as to whether the percentage change was significant, we conducted further statistical tests to address the third objective stated in section 1.2. Thus, we formulated hypothesis one (H1) which states that*: "The use of REVAC parameter tuning during training is similar to the Grid search parameters.".* Table A.3 in the Appendix A shows the outcome of the statistical test conducted for all possible comparisons between the prediction result in Experiment II and Experiment III for each of the three models. The p-values that are bolded and asterisked represents statistically significant difference in performance between models. The lack of significance between models may be due to the small magnitude of the difference in RMSE.

Figure 5.9 compares the performance of the overall predictions (that is, the prediction results over all provinces) made by each model using the original dataset in Experiment I, the upward augmented datasets in Experiment III and the downward augmented dataset in Experiment III. For the CNN model, the overall predictions made with the original, upward augmented, and downward augmented datasets were 31.55, 22.07 and 23.86, respectively. For each dataset, the SVM model's overall RMSE were 33.89, 22.17 and 27.30, respectively while the LSTM model's results were 32.91, 21.93, and 21.60, respectively. These percentages show that using REVAC parameter tuning and various combinations of synthetic and original dataset improves the prediction performance of all three models when compared to making predictions with just the original data. We can also infer that with REVAC parameter tuning, the deep learning models outperformed the SVM model for predictions with both upward and downward dataset augmentation. However, the LSTM model outperformed the CNN model.

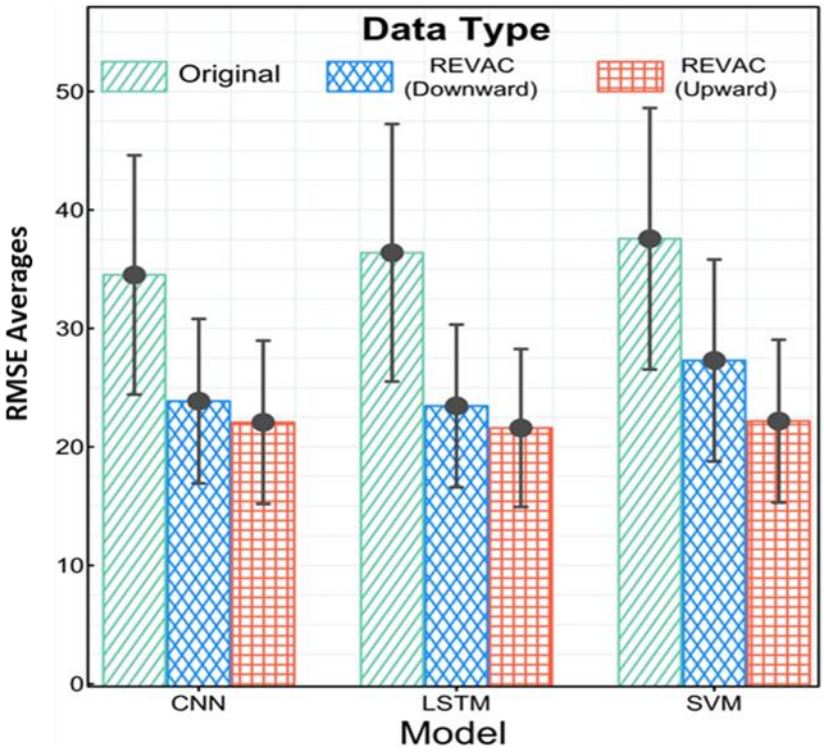

Figure 5. 9: A comparison of CNN, LSTM, SVM average RMSE over South Africa (all provinces) for all prediction scenarios with the original data in Experiment I and all prediction scenarios with the downward augmented data and upward augmented data in Experiment III. Recall that Grid search was used in tuning the parameters of all ML models in Experiment I while REVAC was used in Experiment III. Low RMSE average indicates better performance accuracy. The arrows represent the corresponding widths of twice the standard error. See Table 4.1 for more details on both experiments.

We went further to address the first objective in section 1.2 with Figure 5.10 comparing the average RMSE of the three ML models for both upward and downward augmented dataset predictions conducted in Experiment III for each province. By comparing the models based on the predictions for the upward dataset combinations, we observed that LSTM outperformed the other models in almost every province with the exception of Mpumalanga and Eastern Cape province where CNN recorded lower RMSE for its predictions. The CNN model also outperformed the SVM model in most provinces however, in provinces like Gauteng, Limpopo and KwaZulu Natal, the SVM's prediction results were better than the CNN's. When the models were compared based on their predictions for the downward dataset combinations, LSTM's prediction results still outperformed the other models for most provinces except in provinces like Western Cape, Eastern Cape where CNN's prediction results were better and also in Gauteng province where the SVM's prediction accuracy was higher than the others. The CNN model's predictions also outperformed the SVM's prediction results in most provinces except in provinces like Mpumalanga, Northern Cape, North West and Gauteng where SVM's RMSE were slightly lower.

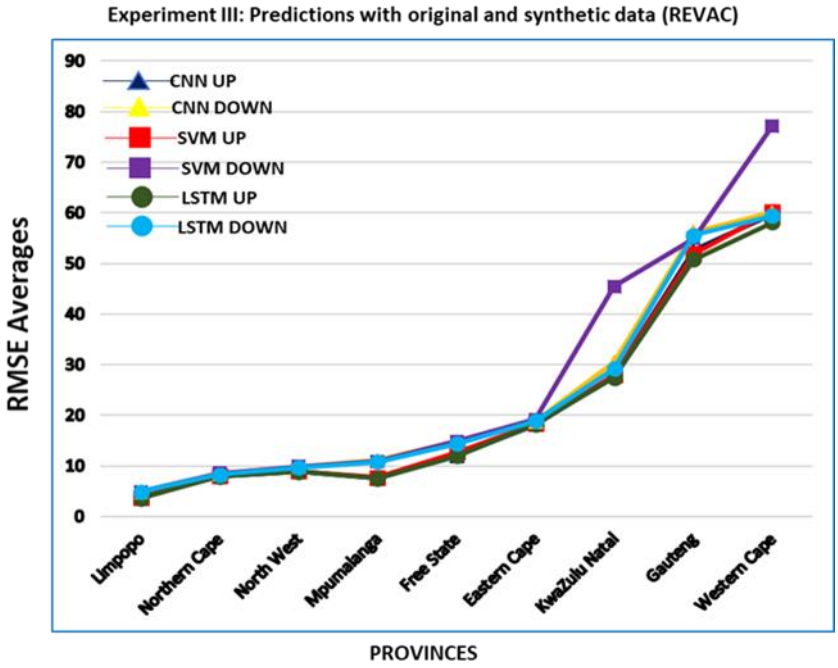

Figure 5.10: *CNN, SVM, LSTM average RMSE error for all prediction scenarios in each province for Experiment III (see Table 4.1 for more details). Lower RMSE averages indicate better performance.*

In addition, the gap in performance margin between the deep learning models and the SVM model were larger in Western Cape and KwaZulu Natal provinces.

To infer whether the difference in performance between the models were statistically significant, we conducted pairwise statistical tests between the three model's RMSE for each province. We used hypothesis zero (H0) as our null hypothesis and applied multiple pairwise Wilcoxon signed rank test with Benjamin Hochberg correction test (see section 4.6 for details). Table A.4 in Appendix A shows the outcome of the statistical test conducted for all possible comparisons between the three model's prediction result in Experiment III for each province and across all provinces. The p-values that are bolded and asterisked represents statistically significant difference in performance between models. The lack of significance between models may be due to the small magnitude of the difference in RMSE. For further details of the analysis carried out in Experiment III see the results in Tables C.1-C.18 in Appendix C.

5.4. Contribution of Climate Factors to the Diarrhoea Prediction Model

Figure 5.11 shows the results of the sensitivity analysis conducted in section 4.7. We observed that the relative importance of each climate variable differs across provinces. For instance, over provinces such as Western Cape, Eastern Cape and Free State, Pressure was the most sensitive to the diarrhoea prediction model. In North West and Mpumalanga, Evaporation was the most important climate variable. In Gauteng, Maximum Temperature was most important while in and KwaZulu Natal, Minimum Temperature was more sensitive. In Limpopo, Humidity was most sensitive variable while Windspeed was more important in Northern Cape.

The most sensitive climate variable in a specific province might be the least sensitive in another. For example, in Eastern Cape, Minimum Temperature was the least sensitive variable meanwhile in KwaZulu Natal, it was most sensitive. In Western Cape, Eastern Cape and Free State, pressure was most important while in Mpumalanga and Northern Cape, it was least important.

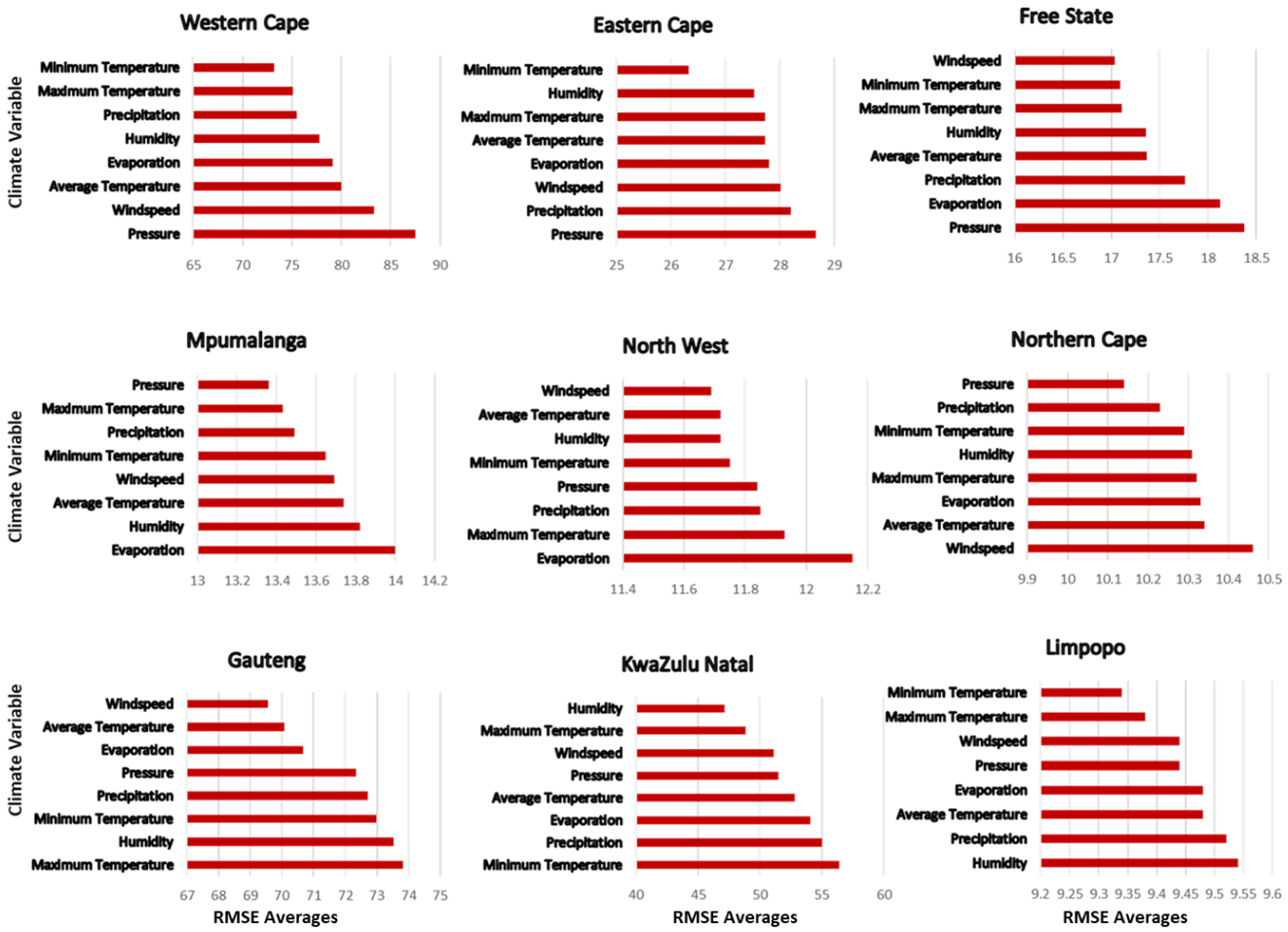

Figure 5. 11: *Variable importance plot for the CNN diarrhoea prediction model in Experiment I for each of 9 South African Province. In each province, the x-axis indicates the prediction accuracy once the variable on the y-axis is omitted from the CNN model. The longer the bar, the larger the loss in accuracy and the higher the importance of that variable in predicting daily diarrhoea cases.*

5.5. Summary of Results

The results showed in this section give details on the outcome of all the experiments conducted for this study. It also relates these outcomes to our research objectives. For instance, observations from the results in section 5.1 can be used to make inference as to which of the three ML models performs best with respect to the original dataset. Results in section 5.2 and 5.3 can be used to make deductions on the influence of data augmentation on the performance of an ML model. Observations in section 5.3 can also be used to make inference on the influence of REVAC parameter tuning on model performance. Finally, the results in section 5.4 can be used to determine the relative importance of each climate variable to diarrhoea outbreak prediction model in each province.

Chapter 6

6. Discussion

This chapter provides a detailed explanation of each ML method's (CNN, LSTM and SVM) performance with respect to the experiments in section 4. We present a detailed analysis of how each ML method performs in different scenarios such as lagging input variables, training with synthetic data, parameter tuning with REVAC evolutionary strategy when compared to one another. We also explain how each climate variable influenced each model's output in each province. Furthermore, we relate our findings to previous studies where Deep learning and traditional ML methods have been used for predicting infectious diseases along with their implications to our research objectives in section 1.2.

6.1. Performance of ML Models for Daily Diarrhoea Case Prediction

The results of all the experiments conducted in section 4 were carried out to test the performance of the three ML methods (SVM, CNN & LSTM) across a range of datasets which was our first objective (see section 1.2 for details). The results showed that all three ML methods were appropriate for predicting daily diarrhoea cases with respect to the selected climate variables in each South African province. Thus, positively validating our first objective. They were all able to yield low and similar RMSE for each prediction scenario in all the given experiments (see Table 4.1 for details on all experiments).

We observed that the RMSE varied across provinces. This is because the RMSE calculates a model's prediction error based on the same unit as the original measurements of the input data which makes interpretation of the error easy to understand [88] . For instance, the input data for Western Cape and Gauteng province were similar thus their resulting RMSE across models are similar.

For each of the experiments conducted, the level of accuracy for each ML model varies (see section 5 for details on all results). Analysis of the results with respect to the different experiment will be given in the following sections.

6.1.1. Performance of Models with the Original Dataset (Experiment I)

To further address our first objective stated above (also in section 1.2), Experiment I was conducted to determine which of the proposed ML methods performed best given the original dataset. In this experiment, predictions were made using only the original dataset.

We can infer from figures 5.1 and 5.2 that the high performance of the deep learning models (that is, CNN and LSTM) may be attributed to the fact that deep learning models are universal approximators and are also able to select important features automatically [8], [12] . In addition, these findings agree with previous research [8], [16], [17], [66] which showed that neural networks and deep learning models outperform traditional ML methods for disease prediction tasks. By observing figure 5.2 and Table A.1 in the appendix, the statistically significant difference in performance between CNN and the LSTM when results are averaged over all provinces may be due to the amount of the dataset used for training used in Experiment I. For instance, [84] showed that LSTM models perform poorly when small datasets size are used for training. Thus, we set up another experiment (Experiment II), to investigate the effect of the size of varying amount of training data on the performance of all models.

6.1.2. Performance of Models with the Augmented Dataset (Experiment II)

Experiment II was conducted to investigate the effect of the augmented training and testing data on model performance in terms of prediction accuracy of the three models which was our second objective. Here, the original dataset was augmented with synthetic data generated by GANs. Observations from the results represented in figures 5.3 and 5.4 show that regardless of how the datasets were augmented the prediction accuracy of all three ML models improved. Over provinces such as Northern Cape, KwaZulu Natal, North West, Mpumalanga, Free State and Western Cape, the performance increase was between 9% and 35% for all models while in other provinces the performance increase was between 35% and 60%. We can also infer from these results that data augmentation boosts model performance in terms of prediction accuracy thus, positively validating the use of data augmentation to improve model accuracy, which was our second objective. In addition, the findings from the prediction tasks conducted in Experiment II are consistent with previous research [21], [22] which demonstrate that data augmentation with synthetic datasets as well as the size of datasets used for training ML models affects the efficiency of their prediction performance.

The use of augmented data improved the task performance of the three models, however, the level of performance increase varied between each ML model. Figures 5.5, 5.6 and Table A.2 in the appendix showed that LSTM statistically significantly outperformed the other ML models in most provinces and across all provinces. This may be because of the increase in the amount of training data used in Experiment II. For instance, studies such as [84] have shown that LSTMs benefit from a large training set size. Another reason may be their ability to easily learn patterns in sequential data. Previous studies like [13], [16] reported that LSTMs are a state of the art for capturing the long-term dependencies specific to a given dataset. When results were averaged over all provinces, the deep learning models significantly outperformed the SVM model. However, the instances such as in Gauteng, Eastern Cape, and Mpumalanga where SVM significantly outperformed the CNN model may be due to the parameter settings of CNN used during training. Note that Grid search parameters were used in both Experiments I and II and previous studies, such as [18], [58] showed that the choice of parameters greatly affects the performance of ML models especially deep learning models. Therefore, we setup a different experiment (Experiment III) where we investigated the effect of REVAC parameter tuning on the performance of all models. The scenarios where there was no significant difference in performance between either of the ML models may be due to the small differences in RMSE between models.

6.1.3. Performance of Models with the Augmented Dataset and REVAC Parameter Tuning (Experiment III)

Experiment III was conducted to investigate to what extent REVAC parameter tuning can improve the accuracy of the three models which was our third objective. REVAC tuning was used to tune the parameters of all three ML models before final predictions were made. Figures 5.7 and 5.8 show that the performance of the CNN model improved over each province by at least 2.5%. Table A. 3 in the appendix also shows that the increase in CNN performance was statistically significant for almost every province. We can infer from these results that REVAC parameter tuning is appropriate for CNNs.

Figures 5.9 and 5.10 showed that with REVAC tuning, the LSTM model still outperformed the other models. Table A.4 in the appendix also show that LSTM statistically significantly outperformed the other models. Even though figures 5.7 and 5.8 showed there was a drop in its performance in provinces like Limpopo, KwaZulu Natal and Free State. Table A.3 in the appendix showed that these decline in performance were not statistically significant. Thus, REVAC parameter tuning may still be appropriate for the LSTM model.

Figure 5.9 showed that with REVAC parameter tuning, SVM model had the least performance across provinces. Figure 5.7 and 5.8 also showed that SVM's performance in Western Cape and KwaZulu Natal, Free State, North West and Mpumalanga provinces

declined by about 1.5%. Table A.3 in the appendix showed that most of the decline in performance were not statistically significant except for Mpumalanga province. The provinces where performance improved were not statistically significant either. Therefore, we can infer from these results that the REVAC parameter tuning is not ideal for the SVM model rather it is more suited to the deep learning models. A possible explanation may be the low dimensional search space of possible parameters for the SVM model considering that an SVM's (with RBF kernel) major parameters are gamma and C only. Studies such as [95] have found that pre-defining a search space for parameter tuning can be difficult. For instance, Nguyen et al. [95] reported that we can miss the optimum parameters if the search space is too large and if it is too small, it may also not contain the optimum parameters. However, [58] reported that grid search is better suited for low dimensional search space perhaps the reason for SVM's satisfactory performance with grid search in Experiments I and II.

The findings from the results in this section shows that REVAC parameter tuning improves the performance of ML learning models. However, the degree of improvement it gives depends on type of the ML algorithm. In this study, we observed that REVAC was better suited to deep learning models. Our findings positively validate the use of REVAC for deep learning algorithms which was our third objective. Our observations are also consistent with previous research [34], [35], [58] that aim to show how the choice of parameter tuning for ML algorithm affects the accuracy of a model.

6.2. Effect of Climate Variables on Diarrhoea Prediction Model (Sensitivity Analysis and Lagged Climate Variables)

By analysing how lagging the climate variables affected the prediction models, Table 5.1 showed that in every province, the models performed better when they are lagged by a certain number of days depending on the selected province. From these findings, we can deduce that antecedent conditions of the climate variables affect the diarrhoea prediction of each province in different ways. This may be attributed to the fact that each province has climate conditions specific to it. Furthermore, a specific climate variable may have a shorter lag effect than another. For instance, Chou et al. showed that [36] temperature variables have a shorter lag effect on diarrhoea than precipitation.

The sensitivity analysis results in section 5.4 provide further evidence as to how differently climate conditions affect a specific province. For instance, over provinces such as Gauteng, KwaZulu Natal, Northern Cape, North West, Western Cape and Mpumalanga, Temperature conditions were among the most sensitive variables to the diarrhoea prediction model. Studies such as [27] have shown that diarrhoea cases increase for every 1°C increase in temperature. While, over provinces such as Mpumalanga, Free State and

North West, evaporation rate was a major contributing climate factor. Kamai et al. [96] reported that evaporation rate is strongly linked to high temperature. Since an increase in diarrhoea cases have been associated with high temperature, perhaps diarrhoea can also be linked to evaporation rate. However, over Limpopo, Mpumalanga, KwaZulu Natal, North West, Eastern Cape and Free State provinces, precipitation and humidity were among the most important variables affecting the diarrhoea prediction model. Several studies such as [8], [36] have also shown that precipitation rate and humidity are strongly related to increase in diarrhoea-related hospitalizations.

From the above findings, we can infer that the contribution of climate factors vary across provinces and that Precipitation, Humidity, Evaporation and Temperature conditions are the most influential factors affecting diarrhoea outbreak in most South African Province.

6.3. Summary and Contributions of Findings

In this section, we discussed the results of the experiments on the performance of three ML methods (CNN, LSTM & SVM) for diarrhoea outbreak prediction over the nine South African provinces with respect to climate factors. The results of the study showed that there was no clear best method overall and that the ML methods possessed different levels of sensitivity to the amount of data available for training and the type of parameter tuning method used during training.

We found that irrespective of the amount of data available for training, the deep learning models outperformed the SVM model. We note that though the real-world data contained fewer data points, the CNN model was able to generalize well and select important features to yield the most satisfactory performance. However, when we augmented the real dataset with synthetic data, the LSTM model outperformed the others. This implies that the LSTM model performs better when the size of training data is large, perhaps the reason for its relatively poor performance in the first experiment.

Although proven useful most studies [8], [16] that use ML methods for diarrhoea outbreak research rely on only real-world dataset. Due to its sensitive nature, health related data is often limited and difficult to access. Therefore, making predictions with the available few might lead to lack of robust conclusions. This study was able to demonstrate the use of both real world and augmented data for diarrhoea outbreak prediction and we found that data augmentation was able to boost the accuracy of all three ML algorithms by over 30% in most provinces.

We also found that the performance of an ML model largely depends on how its parameters are tuned. We adopted two parameter tuning strategies, Grid search and REVAC parameter tuning an evolutionary strategy. While the application of Grid search

tuning has been widely adopted in most ML studies, REVAC parameter tuning on the hand has mainly been used for tuning the parameters of Evolutionary algorithms. This study was able to show that the REVAC algorithm can be adopted for optimizing the performance of ML algorithms especially deep learning models.

Finally, this study provided a foundation for using ML methods to predict diarrhoea outbreak in South Africa. By incorporating climate information, we found that the extent to which climate factors affect the diarrhoea prediction model differs across provinces. Our study has shown that in the prediction of diarrhoea outbreak predictions in South Africa, the most influential climate variables to be considered are precipitation, humidity, evaporation and temperature conditions.

Chapter 7

7. Conclusions

The global burden of diarrhoea cannot be over emphasized as it is a major public health problem that causes both personal and widespread harm. For this reason, we conducted this research to develop a model that could be used for public health surveillance to aid in the prompt notification for the control of diarrhoea outbreak. We compared the performance of three ML methods (CNN, LSTM & SVM) for the prediction of the daily number of diarrhoea cases in the nine South African provinces with respect to eight climate factors in three experiments. The objective of each experiment was to determine which ML method performed best given a specific condition. In the first experiment, predictions were made with only real-world dataset while predictions were made with a combination of real-world and synthetic datasets in the second experiment. In the third experiment, predictions were made with a combination of real and synthetic datasets as well but with REVAC parameter tuning an evolutionary strategy. Our key findings are as follows:

- Overall experiment, the deep learning model's (CNN & LSTM) prediction performance was superior in most provinces. However, the three ML methods possessed different levels of sensitivity to the amount of training data available and the type of parameter tuning method use for training.

- The CNN model performed best when only real-world dataset was used, while the LSTM model outperformed the other models when we augmented the real dataset with synthetic data. However, data augmentation was able to boost the accuracy of all three ML algorithms by over 30% in most provinces.

- The performance of the three ML model improved in most province and across experiments when the datasets were augmented upwards when compared to downward augmentation.

- The performance of an ML model largely depends on how its parameters are tuned. Furthermore, the REVAC algorithm can be adopted for optimizing the performance of ML algorithms especially deep learning models.

- The extent to which climate factors affect the diarrhoea prediction model differs across provinces. For example, in Western Cape and Eastern Cape, some of the most influential climate variables were Pressure and Windspeed while in

Mpumalanga, North West and Northern Cape, Evaporation and Temperature conditions were of major impact.

This work is a preliminary step towards the application of ML methods in the development of an early warning system for predicting the outbreak of diarrhoea in South Africa. We were able to give a deeper understanding on how the amount of data used for training a model can affect the performance of the machine learning model by demonstrating the use of both real-world and augmented data for diarrhoea outbreak prediction. This study was also able to give an insight as to whether the use of the REVAC evolutionary algorithm as a parameter tuning method can improve model performance. Furthermore, the use of Climate-based data for model development further strengthens the claims that climate factors affect diarrhoea.

7.1. Future work

Severe diarrhoea cases that require hospitalizations may not have been taken into considerations. This is because hospital records are difficult to access. This is largely due unavailability of electronic records systems in most clinics and hospitals that use electronic record systems do not give them out due to policies such as confidentiality of patient information. Therefore, the real-world data used in this study may have underestimated the number of diarrhoea cases in a specific province. In addition, due to the few data sources available at our disposal, the real-world diarrhoea dataset we used to conduct our experiments contained a relatively short data-collection period. Another limitation of our study was that designing models for predicting diarrhoea outbreak in South Africa with climate factors alone may be imperfect because the causes of diarrhoea outbreak may involve other human and environmental factors which were not taken into consideration. Furthermore, the variable importance measures provided by the sensitivity analysis may not necessarily indicate the causality of diarrhoea.

Given our current approach of predicting daily number of diarrhoea cases in South Africa, taking other factors that cause the spread of infectious diseases into consideration may improve the accuracy of our diarrhoea prediction model. In addition, gathering real-world data from more than one source such as government databases, public and private hospitals may give a more robust estimate regarding how much diarrhoea cases were recorded in a specific province. Combining data from these sources may also result in a dataset with longer time-period. Finally, given the different strength of each ML algorithm, developing a hybrid model that combines the advantage and benefits of at least two ML algorithms may yield a model that performs better regardless of the conditions set in each experiment.

Bibliography

[1] Guerrant, R. L., Van Gilder, T., Steiner, T. S., Thielman, N. M., Slutsker, L., Tauxe, R. V., ... & Pickering, L. K. (2001). Practice guidelines for the management of infectious diarrhea. *Clinical infectious diseases*, *32*(3), 331-351.

[2] Manatsathit, S., Dupont, H. L., Farthing, M., Kositchaiwat, C., Leelakusolvong, S., Ramakrishna, B. S., ... & Surangsrirat, S. (2002). Guideline for the management of acute diarrhea in adults. *Journal of Gastroenterology and Hepatology*, *17*, S54-S71.

[3] World Health Organization, "The World Health Report 1996," *The World Health Report 1996: Fighting Disease, Fostering Development: Executive Summary,* 1996.

[4] Kosek, M., Bern, C., & Guerrant, R. L. (2003). The global burden of diarrhoeal disease, as estimated from studies published between 1992 and 2000. *Bulletin of the world health organization*, *81*, 197-204.

[5] World Health Organization, "WHO. Protecting Health from Climate Change: World Health Day 2008." In Protecting health from climate change: World Health Day 2008 (pp. 25-25).

[6] Dhimal, M., Karki, K. B., Aryal, K. K., Shrestha, S. L., & Pradhan, B. (2016). Final report on assessment of effects of climate factors on diarrheal diseases at national and sub-national levels in Nepal. *Kathmandu: Nepal Health Research Council and Word Health Organization Country Office for Nepal*.

[7] Jamison, D. T., Breman, J. G., Measham, A. R., Alleyne, G., Claeson, M., Evans, D. B., ... & Musgrove, P. (Eds.). (2006). *Disease control priorities in developing countries*. The World Bank.

[8] Wang, Y., Li, J., Gu, J., Zhou, Z., & Wang, Z. (2015). Artificial neural networks for infectious diarrhea prediction using meteorological factors in Shanghai (China). *Applied Soft Computing*, *35*, 280-290.

[9] Sathya, D., Sudha, V., & Jagadeesan, D. (2020). Application of Machine Learning Techniques in Healthcare. In *Handbook of Research on Applications and Implementations of Machine Learning Techniques* (pp. 289-304). IGI Global.

[10] Sharma, V., Kumar, A., Panat, L., Karajkhede, G., & Lele, A. (2015). Malaria outbreak prediction model using machine learning. *International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)*, *4*(12).

[11] Burke, H. B. (1994, January). Artificial neural networks for cancer research: outcome prediction. In *Seminars in Surgical Oncology* (Vol. 10, No. 1, pp. 73-79). New York: John Wiley & Sons, Inc..

[12] Muniasamy, A., Tabassam, S., Hussain, M. A., Sultana, H., Muniasamy, V., & Bhatnagar, R. (2019, March). Deep learning for predictive analytics in healthcare. In *International Conference on Advanced Machine Learning Technologies and Applications* (pp. 32-42). Springer, Cham.

[13] Pham, T., Tran, T., Phung, D., & Venkatesh, S. (2017). Predicting healthcare trajectories from medical records: A deep learning approach. *Journal of biomedical informatics*, *69*, 218-229.

[14] Tomašev, N., Glorot, X., Rae, J. W., Zielinski, M., Askham, H., Saraiva, A., ... & Mohamed, S. (2019). A clinically applicable approach to continuous prediction of future acute kidney injury. *Nature*, *572*(7767), 116-119.

[15] Dutta, A., Batabyal, T., Basu, M., & Acton, S. T. (2020). An efficient convolutional neural network for coronary heart disease prediction. *Expert Systems with Applications*, *159*, 113408.

[16] Jia, W., Wan, Y., Li, Y., Tan, K., Lei, W., Hu, Y., ... & Xie, G. (2019). Integrating multiple data sources and learning models to predict infectious diseases in China. *AMIA Summits on Translational Science Proceedings*, *2019*, 680.

[17] Hung, C. Y., Chen, W. C., Lai, P. T., Lin, C. H., & Lee, C. C. (2017, July). Comparing deep neural network and other machine learning algorithms for stroke prediction in a large-scale population-based electronic medical claims database. In *2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* (pp. 3110-3113). IEEE.

[18] Vakili, M., Ghamsari, M., & Rezaei, M. (2020). Performance Analysis and Comparison of Machine and Deep Learning Algorithms for IoT Data Classification. *arXiv preprint arXiv:2001.09636*..

[19] Esteban, C., Hyland, S. L., & Rätsch, G. (2017). Real-valued (medical) time series generation with recurrent conditional gans. *arXiv preprint arXiv:1706.02633*.," *arXiv Preprint arXiv:1706.02633,* 2017.

[20] Géron, A. (2019). *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems*. O'Reilly Media..

[21] Fawaz, H. I., Forestier, G., Weber, J., Idoumghar, L., & Muller, P. A. (2018). Data augmentation using synthetic data for time series classification with deep residual networks. *arXiv preprint arXiv:1808.02455*.

[22] Wen, Q., Sun, L., Song, X., Gao, J., Wang, X., & Xu, H. (2020). Time series data augmentation for deep learning: A survey. *arXiv preprint arXiv:2002.12478*.

[23] Bradshaw, D., Groenewald, P., Laubscher, R., Nannan, N., Nojilana, B., Norman, R., ... & Johnson, L. (2003). Initial burden of disease estimates for South Africa, 2000. *South African Medical Journal*, *93*(9), 682-688.

[24] Massyn N, Peer N, English R, Padarath A, Barron P, Day C, editors. District Health Barometer 2015/16. Durban: Health Systems Trust; 2016.

[25] Kullin, B., Meggersee, R., D'Alton, J., Galvao, B., Rajabally, N., Whitelaw, A., ... & Abratt, V. R. (2015). Prevalence of gastrointestinal pathogenic bacteria in patients with diarrhoea attending Groote Schuur Hospital, Cape Town, South Africa. *South African Medical Journal*, *105*(2).

[26] Awotiwon, O. F., Pillay-van Wyk, V., Dhansay, A., Day, C., & Bradshaw, D. (2016). Diarrhoea in children under five years of age in South Africa (1997–2014). *Tropical Medicine & International Health*, *21*(9), 1060-1070.

[27] Musengimana, G., Mukinda, F. K., Machekano, R., & Mahomed, H. (2016). Temperature variability and occurrence of diarrhoea in children under five-years-old in Cape Town metropolitan sub-districts. *International journal of environmental research and public health*, *13*(9), 859.

[28] Davis, C. L., & Vincent, K. (2017). *Climate risk and vulnerability: A handbook for Southern Africa*. CSIR. https://researchspace.csir.co.za/dspace/handle/10204/10066.

[29] Arab, A., Jackson, M. C., & Kongoli, C. (2014). Modelling the effects of weather and climate on malaria distributions in West Africa. *Malaria journal*, *13*(1), 1-9.

[30 Sapankevych, N. I., & Sankar, R. (2009). Time series prediction using support vector machines: a survey. *IEEE Computational Intelligence Magazine*, *4*(2), 24-38.

[31] Olson, R. S., La Cava, W., Orzechowski, P., Urbanowicz, R. J., & Moore, J. H. (2017). PMLB: a large benchmark suite for machine learning evaluation and comparison. *BioData mining*, *10*(1), 1-13.

[32] Caruana, R., & Niculescu-Mizil, A. (2006, June). An empirical comparison of supervised learning algorithms. In *Proceedings of the 23rd international conference on Machine learning* (pp. 161-168).

[33] Probst, P., Bischl, B., & Boulesteix, A. L. (2018). Tunability: Importance of hyperparameters of machine learning algorithms. *arXiv preprint arXiv:1802.09596*.

[34] Nannen, V., & Eiben, A. E. (2007, September). Efficient relevance estimation and value calibration of evolutionary algorithm parameters. In *2007 IEEE congress on evolutionary computation* (pp. 103-110). IEEE.

[35] Smit, S. K., & Eiben, A. E. (2010, July). Beating the 'world champion'evolutionary algorithm via REVAC tuning. In *IEEE Congress on Evolutionary Computation* (pp. 1-8). IEEE.

[36] Chou, W. C., Wu, J. L., Wang, Y. C., Huang, H., Sung, F. C., & Chuang, C. Y. (2010). Modeling the impact of climate variability on diarrhea-associated diseases in Taiwan (1996–2007). *Science of the Total Environment*, *409*(1), 43-51.

[37] Elimian, K. O., Musah, A., Mezue, S., Oyebanji, O., Yennan, S., Jinadu, A., ... & Ihekweazu, C. (2019). Descriptive epidemiology of cholera outbreak in Nigeria, January–November, 2018: implications for the global roadmap strategy. *BMC public health*, *19*(1), 1-11.

[38] Troeger, C., Blacker, B. F., Khalil, I. A., Rao, P. C., Cao, S., Zimsen, S. R., ... & Reiner Jr, R. C. (2018). Estimates of the global, regional, and national morbidity, mortality, and aetiologies of diarrhoea in 195 countries: a systematic analysis for the Global Burden of Disease Study 2016. *The Lancet Infectious Diseases*, *18*(11), 1211-1228.

[39] Kapwata, T., Mathee, A., Le Roux, W. J., & Wright, C. Y. (2018). Diarrhoeal disease in relation to possible household risk factors in South African villages. *International journal of environmental research and public health*, *15*(8), 1665.

[40] Alexander, K. A., Carzolio, M., Goodin, D., & Vance, E. (2013). Climate change is likely to worsen the public health threat of diarrheal disease in Botswana. *International journal of environmental research and public health*, *10*(4), 1202-1230.

[41] Baker, T. (2016). Burden of community diarrhoea in developing countries. *The Lancet. Global Health*, *4*(1), e25.

[42] Azage, M., Kumie, A., Worku, A., C. Bagtzoglou, A., & Anagnostou, E. (2017). Effect of climatic variability on childhood diarrhea and its high risk periods in northwestern parts of Ethiopia. *PLoS One*, *12*(10), e0186933.

[43] Wright, C. Y., Garland, R. M., Norval, M., & Vogel, C. (2014). Human health impacts in a changing South African climate. *South African Medical Journal*, *104*(8), 579-582.

[44] Njidda, A. M., Oyebanji, O., Obasanya, J., Ojo, O., Adedeji, A., Mba, N., ... & Ihekweazu, C. (2018). The Nigeria Centre for Disease Control. *BMJ global health*, *3*(2).

[45] Alexander, K. A., & Blackburn, J. K. (2013). Overcoming barriers in evaluating outbreaks of diarrheal disease in resource poor settings: assessment of recurrent outbreaks in Chobe District, Botswana. *BMC public health*, *13*(1), 1-15.

[46] De Magny, G. C., Murtugudde, R., Sapiano, M. R., Nizam, A., Brown, C. W., Busalacchi, A. J., ... & Colwell, R. R. (2008). Environmental signatures associated with cholera epidemics. *Proceedings of the National Academy of Sciences*, *105*(46), 17676-17681.

[47] Lloyd, S. J., Kovats, R. S., & Armstrong, B. G. (2007). Global diarrhoea morbidity, weather and climate. *Climate Research*, *34*(2), 119-127.

[48] Yan, L., Wang, H., Zhang, X., Li, M. Y., & He, J. (2017). Impact of meteorological factors on the incidence of bacillary dysentery in Beijing, China: A time series analysis (1970-2012). *PLoS One*, *12*(8), e0182937.

[49] McCormick, B. J. J., Alonso, W. J., & Miller, M. A. (2012). An exploration of spatial patterns of seasonal diarrhoeal morbidity in Thailand. *Epidemiology & Infection*, *140*(7), 1236-1243.

[50] Russell, S., & Norvig, P. (2002). Artificial intelligence: a modern approach.

[51] Mitchell, T. M. (1997). Machine learning.

[52] Yu, W., Liu, T., Valdez, R., Gwinn, M., & Khoury, M. J. (2010). Application of support vector machine modeling for prediction of common diseases: the case of diabetes and pre-diabetes. *BMC medical informatics and decision making*, *10*(1), 1-7.

[53] Son, Y. J., Kim, H. G., Kim, E. H., Choi, S., & Lee, S. K. (2010). Application of support vector machine for prediction of medication adherence in heart failure patients. *Healthcare informatics research*, *16*(4), 253-259.

[54] Kilimci, Z. H., Akyuz, A. O., Uysal, M., Akyokus, S., Uysal, M. O., Atak Bulbul, B., & Ekmis, M. A. (2019). An improved demand forecasting model using deep learning approach and proposed decision integration strategy for supply chain. *Complexity*, *2019*.

[55] Nguyen, D., Nguyen, C., Duong-Ba, T., Nguyen, H., Nguyen, A., & Tran, T. (2017, January). Joint network coding and machine learning for error-prone wireless broadcast. In *2017 IEEE 7th Annual Computing and Communication Workshop and Conference (CCWC)* (pp. 1-7). IEEE.

[56] Brabazon, A., O'Neill, M., & McGarraghy, S. (2015). Neural Networks for Supervised Learning. In *Natural Computing Algorithms* (pp. 221-259). Springer, Berlin, Heidelberg.

[57] Al-Shayea, Q. K. (2011). Artificial neural networks in medical diagnosis. *International Journal of Computer Science Issues*, *8*(2), 150-154.

[58] J. Wu, J., Chen, X. Y., Zhang, H., Xiong, L. D., Lei, H., & Deng, S. H. (2019). Hyperparameter optimization for machine learning models based on Bayesian optimization. *Journal of Electronic Science and Technology*, *17*(1), 26-40.

[59] Lundervold, A. S., & Lundervold, A. (2019). An overview of deep learning in medical imaging focusing on MRI. *Zeitschrift für Medizinische Physik*, *29*(2), 102-127.

[60] LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *nature*, *521*(7553), 436-444.

[61] Kiranyaz, S., Avci, O., Abdeljaber, O., Ince, T., Gabbouj, M., & Inman, D. J. (2021). 1D convolutional neural networks and applications: A survey. *Mechanical Systems and Signal Processing*, *151*, 107398.

[62] Huang, C. J., & Kuo, P. H. (2019). Multiple-input deep convolutional neural network model for short-term photovoltaic power forecasting. *IEEE Access*, *7*, 74822-74834.

[63] Helmini, S., Jihan, N., Jayasinghe, M., & Perera, S. (2019). Sales forecasting using multivariate long short term memory network models. *PeerJ PrePrints*, *7*, e27712v1.

[64] Adamker, G., Holzer, T., Karakis, I., Amitay, M., Anis, E., Singer, S. R., & Barnett-Itzhaki, Z. (2018). Prediction of Shigellosis outcomes in Israel using machine learning classifiers. *Epidemiology & Infection*, *146*(11), 1445-1451.

[65] Akbar, W., Wu, W. P., Saleem, S., Farhan, M., Saleem, M. A., Javeed, A., & Ali, L. (2020). Development of Hepatitis Disease Detection System by Exploiting Sparsity in Linear Support Vector Machine to Improve Strength of AdaBoost Ensemble Model. *Mobile Information Systems*, *2020*.

[66] Chae, S., Kwon, S., & Lee, D. (2018). Predicting infectious disease using deep learning and big data. *International journal of environmental research and public health*, *15*(8), 1596.

[67] Abideen, Zain Ul, Mubeen Ghafoor, Kamran Munir, Madeeha Saqib, Ata Ullah, Tehseen Zia, Syed Ali Tariq, Ghufran Ahmed, and Asma Zahra. "Uncertainty assisted robust tuberculosis identification with bayesian convolutional neural networks." *Ieee Access* 8 (2020): 22812-22825.

[68] Fuhad, K. M., Tuba, J. F., Sarker, M., Ali, R., Momen, S., Mohammed, N., & Rahman, T. (2020). Deep learning based automatic malaria parasite detection from blood smear and its smartphone based application. *Diagnostics*, *10*(5), 329.

[69] Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, *15*(1), 1929-1958.

[70] Fang, X., Liu, W., Ai, J., He, M., Wu, Y., Shi, Y., ... & Bao, C. (2020). Forecasting incidence of infectious diarrhea using random forest in Jiangsu Province, China. *BMC infectious diseases*, *20*(1), 1-8.

[71] *WHO. Coronavirus disease (COVID-19) outbreak situation; 2020*. Available: https://www.who.int/emergencies/diseases/novel-coronavirus-2019.

[72] Senior, A. W., Evans, R., Jumper, J., Kirkpatrick, J., Sifre, L., Green, T., ... & Hassabis, D. (2020). Improved protein structure prediction using potentials from deep learning. *Nature*, *577*(7792), 706-710.

[73] Bullock, J., Luccioni, A., Pham, K. H., Lam, C. S. N., & Luengo-Oroz, M. (2020). Mapping the landscape of artificial intelligence applications against COVID-19. *Journal of Artificial Intelligence Research*, *69*, 807-845.

[74] Hu, F., Jiang, J., & Yin, P. (2020). Prediction of potential commercially inhibitors against SARS-CoV-2 by multi-task deep model. *arXiv preprint arXiv:2003.00728*.

[75] Beck, B. R., Shin, B., Choi, Y., Park, S., & Kang, K. (2020). Predicting commercially available antiviral drugs that may act on the novel coronavirus (SARS-CoV-2) through a drug-target interaction deep learning model. *Computational and structural biotechnology journal*, *18*, 784-790.

[76] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

[77] Mohapatra, S., Nath, P., Chatterjee, M., Das, N., Kalita, D., Roy, P., & Satapathi, S. (2020). Repurposing therapeutics for COVID-19: rapid prediction of commercially available drugs through machine learning and docking. *Plos one*, *15*(11), e0241543.

[78] Zhavoronkov, A., Aladinskiy, V., Zhebrak, A., Zagribelnyy, B., Terentiev, V., Bezrukov, D. S., ... & Ivanenkov, Y. (2020). Potential COVID-2019 3C-like protease inhibitors designed using generative deep learning approaches. *Insilico Medicine Hong Kong Ltd A*, *307*, E1.

[79] Ong, E., Wong, M. U., Huffman, A., & He, Y. (2020). COVID-19 coronavirus vaccine design using reverse vaccinology and machine learning. *Frontiers in immunology*, *11*, 1581.

[80] Bandyopadhyay, S. K., & Dutta, S. (2020). Machine learning approach for confirmation of covid-19 cases: Positive, negative, death and release. *medRxiv*.

[81] Ye, Yanfang, Shifu Hou, Yujie Fan, Yiyue Qian, Yiming Zhang, Shiyu Sun, Qian Peng, and Kenneth Laparo. "$\alpha $-Satellite: An AI-driven System and Benchmark Datasets for Hierarchical Community-level Risk Assessment to Help Combat COVID-19." *arXiv preprint arXiv:2003.12232* (2020).

[82] Lopez-Rincon, A., Tonda, A., Mendoza-Maldonado, L., Claassen, E., Garssen, J., & Kraneveld, A. D. (2020). Accurate identification of sars-cov-2 from viral genome sequences using deep learning. *bioRxiv*.

[83] O'Mahony, N., Campbell, S., Carvalho, A., Harapanahalli, S., Hernandez, G. V., Krpalkova, L., ... & Walsh, J. (2019, April). Deep learning vs. traditional computer vision. In *Science and Information Conference* (pp. 128-144). Springer, Cham.

[84] Yang, S., Yu, X., & Zhou, Y. (2020, June). LSTM and GRU Neural Network Performance Comparison Study: Taking Yelp Review Dataset as an Example. In *2020 International Workshop on Electronic Communication and Artificial Intelligence (IWECAI)* (pp. 98-101). IEEE.

[85] Karegowda, A. G., Manjunath, A. S., & Jayaram, M. A. (2011). Application of genetic algorithm optimized neural network connection weights for medical diagnosis of pima Indians diabetes. *International Journal on Soft Computing*, *2*(2), 15-23.

[86] Kalnay, E., Kanamitsu, M., Kistler, R., Collins, W., Deaven, D., Gandin, L., ... & Joseph, D. (1996). The NCEP/NCAR 40-year reanalysis project. *Bulletin of the American meteorological Society*, *77*(3), 437-472.

[87] Chai, T., & Draxler, R. R. (2014). Root mean square error (RMSE) or mean absolute error (MAE)?–Arguments against avoiding RMSE in the literature. *Geoscientific model development*, *7*(3), 1247-1250.

[88] Pelánek, R. (2015). Metrics for Evaluation of Student Models. *Journal of Educational Data Mining*, *7*(2), 1-19.

[89] Shapiro, S. S., & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, *52*(3/4), 591-611.

[90] Wilcoxon, F. (1992). Individual comparisons by ranking methods. In *Breakthroughs in statistics* (pp. 196-202). Springer, New York, NY.

[91] Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *The Journal of Machine Learning Research*, *7*, 1-30.

[92] Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *The Journal of Machine Learning Research*, *7*, 1-30.

[93] Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, *57*(1), 289-300.

[94] Gevrey, M., Dimopoulos, I., & Lek, S. (2003). Review and comparison of methods to study the contribution of variables in artificial neural network models. *Ecological modelling*, *160*(3), 249-264.

[95] Nguyen, V., Gupta, S., Rane, S., Li, C., & Venkatesh, S. (2017, November). Bayesian optimization in weakly specified search space. In *2017 IEEE International Conference on Data Mining (ICDM)* (pp. 347-356). IEEE.

[96] Kamai, T., Weisbrod, N., & Dragila, M. I. (2009). Impact of ambient temperature on evaporation from surface-exposed fractures. *Water resources research*, *45*(2).

[97] Choi, E., Biswal, S., Malin, B., Duke, J., Stewart, W. F., & Sun, J. (2017, November). Generating multi-label discrete patient records using generative adversarial networks. In *Machine learning for healthcare conference* (pp. 286-305). PMLR.

[98] Che, Z., Cheng, Y., Zhai, S., Sun, Z., & Liu, Y. (2017, November). Boosting deep learning risk prediction with generative adversarial networks for electronic health records. In *2017 IEEE International Conference on Data Mining (ICDM)* (pp. 787-792). IEEE.

[99] Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... & Bengio, Y. (2014). Generative adversarial networks. *arXiv preprint arXiv:1406.2661*.

[100] Hayes, J., Melis, L., Danezis, G., & De Cristofaro, E. (2017). Logan: Membership inference attacks against generative models. *arXiv preprint arXiv:1705.07663*.

Appendices

Appendix A

P-values for Experiments I, II & III

This appendix section contains the outcome of the pairwise Wilcoxon signed rank test conducted between models in Experiments I, II & III. Experiments I involved predictions with original data only while Experiments II & III involved predictions with combinations of original and synthetic data generated by GANS. However, in Experiment III REVAC parameter tuning was used to tune the parameters of all ML models. See Table 4.1 for more details on each experiment.

Table A. 1: Wilcoxon signed rank test Adjusted p-values for the pair-wise comparisons of the three ML methods within province based on the average RMSE errors for all prediction scenarios the with original data in Experiment I. Recall that Grid search was used to tune the parameters of all ML models in Experiment I (see Table 4.1 for details). H_0 is the null hypothesis while Ha represent the alternate hypothesis. * represents p-values that are statistically significant. Please note that H_0 (No statistical significance) indicates that the performance between ML method 1 and ML method 2 are similar while Ha (Statistically significant difference) indicates that the model with smaller RMSE is significantly better than the other (see section 4.6 for details).

ML Method1 - ML Method2
H_0: ML Method1 = ML Method2
Ha: ML Method1 ≠ ML Method2

LSTM - CNN									
Eastern Cape	Free State	Gauteng	KwaZulu Natal	Limpopo	Mpumalanga	Northern Cape	North West	Western Cape	Average of Provinces
0.8705	0.3409	0.2813	0.4688	0.1974	0.1442	0.1974	0.1442	0.1974	**0.0032***
LSTM - SVM									
Eastern Cape	Free State	Gauteng	KwaZulu Natal	Limpopo	Mpumalanga	Northern Cape	North West	Western Cape	Average of Provinces
0.1442	1.0000	0.7212	0.7212	0.1974	0.1442	0.1442	0.4688	1.0000	0.1442
SVM - CNN									
Eastern Cape	Free State	Gauteng	KwaZulu Natal	Limpopo	Mpumalanga	Northern Cape	North West	Western Cape	Average of Provinces
0.1442	0.3409	0.1442	0.1974	0.1974	0.1442	0.1442	0.1442	0.1974	**8.13e-07***

Table A. 2 Wilcoxon signed rank test adjusted p-values for the pair-wise comparisons of the three ML methods within province based on the average RMSE errors for all prediction scenarios with combinations of synthetic and original data (augmented both upwards and downward) in Experiment II. Recall that the parameters from the Grid search in Experiment I were maintained in this experiment (see Table 4.1 for details). H_0 is the null hypothesis while Ha represent the alternate hypothesis. * represents p-values that are statistically significant. Please note that H_0 (No statistical significance) indicates that the performance between ML method 1 and ML method 2 are similar while Ha (Statistically significant difference) indicates that the model with smaller RMSE is significantly better than the other (see section 4.6 for details).

ML Method1 – ML Method2
H_0: ML Method1 = ML Method2
Ha: ML Method1 ≠ ML Method2

LSTM – CNN (Upward augmentation)									
Eastern Cape	Free State	Gauteng	KwaZulu Natal	Limpopo	Mpumalanga	Northern Cape	North West	Western Cape	Average of Provinces
1.29e-01	2.39e-01	1.39e-01	6.66e-02	**5.35e-03***	2.53e-01	**2.83e-02***	**5.31e-03***	**3.96e-02***	**7.12e-08***
LSTM – SVM (Upward augmentation)									
Eastern Cape	Free State	Gauteng	KwaZulu Natal	Limpopo	Mpumalanga	Northern Cape	North West	Western Cape	Average of Provinces
2.26e-01	2.26e-01	8.06e-01	**1.69e-02***	6.81e-02	**3.99e-02***	**2.43e-02***	**3.68e-02***	9.52e-01	**5.31e-03***
SVM – CNN (Upward augmentation)									
Eastern Cape	Free State	Gauteng	KwaZulu Natal	Limpopo	Mpumalanga	Northern Cape	North West	Western Cape	Average of Provinces
2.43e-02*	3.20e-01	1.39e-01	6.17e-01	1.39e-01	4.30e-01	**3.29e-02***	**2.43e-02***	1.39e-01	**8.23e-06***
LSTM – CNN (Downward augmentation)									
Eastern Cape	Free State	Gauteng	KwaZulu Natal	Limpopo	Mpumalanga	Northern Cape	North West	Western Cape	Average of Provinces
7.61e-03*	1.08e-01	**1.70e-03***	**2.29e-05***	**8.34e-04***	**1.43e-05***	**8.37e-04***	**2.45e-05***	**7.73e-02***	**6.60e-15***
LSTM – SVM (Downward augmentation)									
Eastern Cape	Free State	Gauteng	KwaZulu Natal	Limpopo	Mpumalanga	Northern Cape	North West	Western Cape	Average of Provinces
5.88e-02	**3.04e-04***	**2.45e-02***	**2.45e-05***	7.47e-01	**5.72e-03***	**2.97e-03***	**2.47e-02***	**3.34e-04***	**8.76e-07***
SVM – CNN (Downward augmentation)									
Eastern Cape	Free State	Gauteng	KwaZulu Natal	Limpopo	Mpumalanga	Northern Cape	North West	Western Cape	Average of Provinces
1.39e-02*	**1.43e-05***	**1.10e-03***	**7.44e-04***	**2.97e-03***	**2.86e-04***	**1.37e-03***	**2.10e-04***	**3.58e-05***	8.95e-02

Table A. 3: Wilcoxon signed rank test adjusted p-values for the pair-wise comparisons of the REVAC tuning method in Experiment III and the Grid search parameters in Experiment II for the three ML methods. Each comparison was within province and based on the average RMSE errors for all prediction scenarios with the combination of synthetic and original data augmented both upwards and downward in Experiment III and Experiment II (see Table 4.1 for details on both experiments). H_1 is the null hypothesis while Ha represent the alternate hypothesis. * represents p-values that are statistically significant. Please note that H_0 (No statistical significance) indicates that the performance between tuning method 1 and tuning method 2 are similar while Ha (Statistically significant difference) indicates that the tuning method that yields a smaller RMSE is significantly better than the other (see section 4.6 for details).

Tuning Method1 - Tuning Method2
H_1: Tuning Method1 = Tuning Method2
H_a: Tuning Method1 ≠ Tuning Method2

CNN: Grid vs REVAC (Upward)									
Eastern Cape	Free State	Gauteng	KwaZulu Natal	Limpopo	Mpumalanga	Northern Cape	North West	Western Cape	Average of Provinces
2.86e-05*	**8.32e-03***	1.86e-01	2.38e-01	3.62e-01	**5.98e-03***	**1.02e-03***	**3.62e-04***	1.12e-01	**7.15e-12***
LSTM: Grid vs REVAC (Upward)									
Eastern Cape	Free State	Gauteng	KwaZulu Natal	Limpopo	Mpumalanga	Northern Cape	North West	Western Cape	Average of Provinces
2.48e-01	**2.26e-02***	**2.31e-02***	3.07e-01	3.19e-01	1.83e-01	**2.26e-02***	3.07e-01	**2.31e-02***	**1.02e-03***
SVM: Grid vs REVAC (Upward)									
Eastern Cape	Free State	Gauteng	KwaZulu Natal	Limpopo	Mpumalanga	Northern Cape	North West	Western Cape	Average of Provinces
2.46e-01	3.62e-01	7.56e-02	5.97e-02	7.85e-01	7.85e-01	9.26e-01	9.26e-01	1.65e-01	3.19e-01
CNN: Grid vs REVAC (Downward)									
Eastern Cape	Free State	Gauteng	KwaZulu Natal	Limpopo	Mpumalanga	Northern Cape	North West	Western Cape	Average of Provinces
6.45e-04*	**5.08e-03***	**6.99e-04***	**6.45e-04***	**6.45e-04***	**1.43e-04***	**1.09e-03***	**6.45e-04***	**6.99e-04***	**6.60e-15***
LSTM: Grid vs REVAC (Downward)									
Eastern Cape	Free State	Gauteng	KwaZulu Natal	Limpopo	Mpumalanga	Northern Cape	North West	Western Cape	Average of Provinces
5.55e-01	8.40e-01	8.70e-01	8.70e-01	**4.68e-03***	1.46e-01	2.34e-01	3.14e-01	**4.97e-03***	**2.66e-02***
SVM: Grid vs REVAC (Downward)									
Eastern Cape	Free State	Gauteng	KwaZulu Natal	Limpopo	Mpumalanga	Northern Cape	North West	Western Cape	Average of Provinces
8.40e-01	9.97e-02	8.40e-01	3.04e-01	7.03e-01	**6.45e-04***	2.82e-01	4.96e-01	8.81e-02	4.44e-01

Table A. 4: Wilcoxon signed rank test adjusted p-values for the pair-wise comparisons of the three ML methods within province based on the average RMSE errors for all prediction scenarios with combinations of synthetic and original data (augmented both upwards and downward) in Experiment III. Recall that REVAC was used to tune parameters of all ML models in this experiment (see Table 4.1 for details). H_0 is the null hypothesis while Ha represent the alternate hypothesis. * represents p-values that are statistically significant. Please note that H_0 (No statistical significance) indicates that the performance between ML method 1 and ML method 2 are similar while Ha (Statistically significant difference) indicates that the model with smaller RMSE is significantly better than the other (see section 4.6 for details).

ML Method1 - ML Method2
H_0: ML Method1 = ML Method2
Ha: ML Method1 ≠ ML Method2

Eastern Cape	Free State	Gauteng	KwaZulu Natal	Limpopo	Mpumalanga	Northern Cape	North West	Western Cape	Average of Provinces
LSTM – CNN (Upward augmentation & REVAC)									
4.40e-01	5.20e-01	**1.91e-05***	4.40e-01	**1.26e-03***	5.20e-01	8.96e-01	1.09e-01	**2.90e-03***	**1.35e-05***
LSTM – SVM (Upward augmentation & REVAC)									
4.55e-01	**6.00e-03***	2.02e-01	8.40e-01	**5.96e-03***	2.00e-01	1.15e-01	**1.70e-02***	**5.96e-03***	**6.62e-09***
SVM – CNN (Upward augmentation & REVAC)									
4.40e-01	**1.26e-03***	4.38e-01	8.40e-01	**2.18e-02***	5.90e-02	**6.39e-03***	8.40e-01	8.96e-01	**2.90e-03***
LSTM – CNN (Downward augmentation & REVAC)									
3.62e-01	5.51e-02	2.70e-01	**2.87e-02***	1.64e-01	**1.33e-02***	**1.27e-02***	5.51e-02	5.51e-02	**1.35e-05***
LSTM – SVM (Downward augmentation & REVAC)									
2.89e-02*	**4.48e-02***	**1.96e-02***	**1.43e-05***	1.14e-01	4.22e-01	**2.01e-02***	1.14e-01	**9.54e-05***	**6.62e-09***
SVM – CNN (Downward augmentation & REVAC)									
1.36e-02*	8.41e-01	**2.57e-02***	**1.43e-05***	7.70e-01	1.22e-01	3.01e-01	4.26e-01	**9.54e-05***	**2.90e-03***

Appendix B

RMSE for all Prediction Scenarios in Experiment II (predictions with combinations of original and synthetic data)

Table B. 1: *RMSE errors from the CNN, SVM and LSTM models for all Western Cape dataset combinations mixed upwards for predictions in Experiment II. The total synthetic to original data ratio is used in the proportions shown in the table. Lower RMSE percentages indicate better prediction accuracy of the model and vice-versa*.

Western Cape Dataset Synthetic: Original (20,000:3763)	***CNN RMSE when mixed upwards (Lag1)***	***SVM RMSE when mixed upwards (Lag1)***	***LSTM RMSE when mixed upwards (Lag1)***	***CNN RMSE when mixed upwards (Lag5)***	***SVM RMSE when mixed upwards (Lag5)***	***LSTM RMSE when mixed upwards (Lag5)***	***CNN RMSE when mixed upwards (Lag14)***	***SVM RMSE when mixed upwards (Lag14)***	***LSTM RMSE when mixed upwards (Lag14)***	***CNN RMSE when mixed upwards (Lag21)***	***SVM RMSE when mixed upwards (Lag21)***	***LSTM RMSE when mixed upwards (Lag21)***
90/10 percent	69.16	72.64	69.04	48.02	46.47	46.55	58.30	55.83	56.15	70.21	58.81	60.55
80/20 percent	69.62	73.62	71.35	48.47	46.81	47.40	58.35	55.74	56.28	70.30	58.95	61.23
70/30 percent	69.51	74.03	72.36	51.83	47.37	47.37	57.43	55.64	56.87	67.81	59.27	59.62
60/40 percent	70.06	77.04	74.59	49.76	48.58	48.71	58.67	55.22	54.78	76.16	59.73	62.11
50/50 percent	72.05	80.38	77.55	53.59	49.02	49.63	60.50	56.54	54.87	71.80	60.87	60.90

Table B.2: RMSE errors from the CNN, SVM and LSTM models for all Western Cape dataset combinations mixed downwards for predictions in Experiment II. The total synthetic to original data ratio is used in the proportions shown in the table. Lower RMSE percentages indicate better prediction accuracy of the model and vice-versa.

Western Cape Dataset Synthetic: Original (20,000:3763)	***CNN RMSE when mixed downwards (Lag1)***	***SVM RMSE when mixed downwards (Lag 1)***	***LSTM RMSE when mixed downwards (Lag 1)***	***CNN RMSE when mixed downwards (Lag5)***	***SVM RMSE when mixed downwards (Lag5)***	***LSTM RMSE when mixed downwards (Lag5)***	***CNN RMSE when mixed downwards (Lag14)***	***SVM RMSE when mixed downwards (Lag14)***	***LSTM RMSE when mixed downwards (Lag14)***	***CNN RMSE when mixed downwards (Lag21)***	***SVM RMSE when mixed downwards (Lag21)***	***LSTM RMSE when mixed downwards (Lag21)***
90/10 percent	68.32	72.07	68.59	48.65	50.01	45.28	57.06	78.79	55.79	69.65	77.97	58.86
80/20 percent	69.23	71.52	70.14	48.17	54.06	48.53	59.74	95.85	54.57	76.86	98.08	58.25
70/30 percent	69.56	71.53	71.51	51.07	56.20	50.04	55.56	111.45	55.64	68.61	99.56	61.15
60/40 percent	75.16	74.24	76.39	50.45	59.40	51.01	60.32	122.51	56.26	79.10	104.13	58.73
50/50 percent	83.10	81.87	94.61	59.76	75.84	57.56	68.09	114.82	68.19	82.45	108.06	62.67

Table B.3: RMSE errors from the CNN, SVM and LSTM models for all KwaZulu Natal dataset combinations mixed upwards for predictions in Experiment II. The total synthetic to original data ratio is used in the proportions shown in the table. Lower RMSE percentages indicate better prediction accuracy of the model and vice-versa.

KwaZulu Natal Dataset Synthetic: Original (20,000:3763)	***CNN RMSE when mixed upwards (Lag1)***	***SVM RMSE when mixed upwards (Lag1)***	***LSTM RMSE when mixed upwards (Lag1)***	***CNN RMSE when mixed upwards (Lag5)***	***SVM RMSE when mixed upwards (Lag5)***	***LSTM RMSE when mixed upwards (Lag5)***	***CNN RMSE when mixed upwards (Lag14)***	***SVM RMSE when mixed upwards (Lag14)***	***LSTM RMSE when mixed upwards (Lag14)***	***CNN RMSE when mixed upwards (Lag21)***	***SVM RMSE when mixed upwards (Lag21)***	***LSTM RMSE when mixed upwards (Lag21)***
90/10 percent	43.14	48.35	44.23	20.47	21.28	19.44	22.71	20.10	19.00	27.76	21.13	20.44
80/20 percent	41.68	47.09	44.78	20.50	21.41	20.29	21.96	20.18	18.86	25.62	20.90	21.43
70/30 percent	44.24	47.93	45.33	22.67	21.82	21.56	23.30	20.28	19.50	25.91	20.93	21.61
60/40 percent	43.49	49.50	46.92	22.34	21.78	20.76	22.88	21.15	20.74	26.71	21.48	22.01
50/50 percent	45.21	50.90	48.53	21.71	22.27	21.42	23.69	22.28	21.39	26.55	21.88	23.52

Table B.4: RMSE errors from the CNN, SVM and LSTM models for all KwaZulu Natal dataset combinations mixed downwards for predictions in Experiment II. The total synthetic to original data ratio is used in the proportions shown in the table. Lower RMSE percentages indicate better prediction accuracy of the model and vice-versa.

KwaZulu Natal Dataset Synthetic: Original (20,000:3763)	***CNN RMSE when mixed downwards (Lag1)***	***SVM RMSE when mixed downwards (Lag 1)***	***LSTM RMSE when mixed downwards (Lag 1)***	***CNN RMSE when mixed downwards (Lag5)***	***SVM RMSE when mixed downwards (Lag5)***	***LSTM RMSE when mixed downwards (Lag5)***	***CNN RMSE when mixed downwards (Lag14)***	***SVM RMSE when mixed downwards (Lag14)***	***LSTM RMSE when mixed downwards (Lag14)***	***CNN RMSE when mixed downwards (Lag21)***	***SVM RMSE when mixed downwards (Lag21)***	***LSTM RMSE when mixed downwards (Lag21)***
90/10 percent	43.70	49.26	44.29	25.89	28.66	20.01	22.11	33.36	19.56	26.07	33.28	20.11
80/20 percent	45.95	48.97	45.39	23.38	32.74	21.59	28.97	41.87	19.44	26.59	41.48	21.33
70/30 percent	45.44	49.06	46.39	28.59	35.29	24.09	29.12	46.46	20.38	30.90	46.69	21.43
60/40 percent	45.93	48.86	48.02	33.06	36.21	27.22	32.50	52.59	22.30	41.84	52.17	23.55
50/50 percent	52.24	51.99	52.96	38.15	46.11	36.55	34.21	60.76	28.06	48.89	60.09	28.29

Table *B.5: RMSE errors from the CNN, SVM and LSTM models for all Limpopo dataset combinations mixed upwards for predictions in Experiment II. The total synthetic to original data ratio is used in the proportions shown in the table. Lower RMSE percentages indicate better prediction accuracy of the model and vice-versa.*

Limpopo Dataset ***Synthetic: Original (20,000:3763)***	***CNN RMSE when mixed upwards (Lag1)***	***SVM RMSE when mixed upwards (Lag1)***	***LSTM RMSE when mixed upwards (Lag1)***	***CNN RMSE when mixed upwards (Lag5)***	***SVM RMSE when mixed upwards (Lag5)***	***LSTM RMSE when mixed upwards (Lag5)***	***CNN RMSE when mixed upwards (Lag14)***	***SVM RMSE when mixed upwards (Lag14)***	***LSTM RMSE when mixed upwards (Lag14)***	***CNN RMSE when mixed upwards (Lag21)***	***SVM RMSE when mixed upwards (Lag21)***	***LSTM RMSE when mixed upwards (Lag21)***
90/10 percent	4.54	4.86	3.34	2.99	3.22	3.02	3.57	3.35	3.34	3.95	3.38	3.35
80/20 percent	4.57	4.80	3.47	3.04	3.12	2.92	3.50	3.33	3.47	4.21	3.35	3.40
70/30 percent	4.47	4.85	3.39	3.20	3.13	2.96	3.78	3.38	3.52	4.03	3.33	4.77
60/40 percent	4.69	4.97	3.50	3.24	3.19	2.95	3.83	3.42	3.50	3.98	3.42	3.41
50/50 percent	4.80	5.07	3.44	3.40	3.19	2.95	3.89	3.43	3.44	4.02	3.51	3.51

Table B.6: RMSE errors from the CNN, SVM and LSTM models for all Limpopo dataset combinations mixed downwards for predictions in Experiment II. The total synthetic to original data ratio is used in the proportions shown in the table. Lower RMSE percentages indicate better prediction accuracy of the model and vice-versa.

Limpopo Dataset ***Synthetic: Original (20,000:3763)***	***CNN RMSE when mixed downwards (Lag1)***	***SVM RMSE when mixed downwards (Lag 1)***	***LSTM RMSE when mixed downwards (Lag 1)***	***CNN RMSE when mixed downwards (Lag5)***	***SVM RMSE when mixed downwards (Lag5)***	***LSTM RMSE when mixed downwards (Lag5)***	***CNN RMSE when mixed downwards (Lag14)***	***SVM RMSE when mixed downwards (Lag14)***	***LSTM RMSE when mixed downwards (Lag14)***	***CNN RMSE when mixed downwards (Lag21)***	***SVM RMSE when mixed downwards (Lag21)***	***LSTM RMSE when mixed downwards (Lag21)***
90/10 percent	4.77	4.99	4.78	3.42	3.38	3.27	4.00	3.63	3.80	4.42	3.62	3.61
80/20 percent	4.91	5.07	4.99	3.72	3.73	3.68	4.34	3.89	4.09	4.68	3.92	3.88
70/30 percent	5.34	5.18	5.07	4.35	4.17	4.18	4.67	4.33	4.45	4.93	4.33	4.29
60/40 percent	5.77	5.83	5.80	5.30	5.01	5.26	5.57	5.16	5.35	5.90	5.41	5.42
50/50 percent	7.08	7.17	7.05	6.70	6.42	6.54	6.98	6.53	6.58	7.32	6.71	6.71

Table B.7: RMSE errors from the CNN, SVM and LSTM models for all Free State dataset combinations mixed upwards for predictions in Experiment II. The total synthetic to original data ratio is used in the proportions shown in the table. Lower RMSE percentages indicate better prediction accuracy of the model and vice-versa.

Free State Dataset Synthetic: Original (20,000:3763)	***CNN RMSE when mixed upwards (Lag1)***	***SVM RMSE when mixed upwards (Lag1)***	***LSTM RMSE when mixed upwards (Lag1)***	***CNN RMSE when mixed upwards (Lag5)***	***SVM RMSE when mixed upwards (Lag5)***	***LSTM RMSE when mixed upwards (Lag5)***	***CNN RMSE when mixed upwards (Lag14)***	***SVM RMSE when mixed upwards (Lag14)***	***LSTM RMSE when mixed upwards (Lag14)***	***CNN RMSE when mixed upwards (Lag21)***	***SVM RMSE when mixed upwards (Lag21)***	***LSTM RMSE when mixed upwards (Lag21)***
90/10 percent	14.35	15.68	15.45	8.52	8.71	8.42	12.29	11.59	11.24	15.05	13.14	12.53
80/20 percent	14.48	15.70	15.95	8.77	8.87	9.01	13.52	11.51	11.09	16.76	13.39	13.36
70/30 percent	14.30	15.53	16.10	8.75	8.83	8.79	12.22	11.77	11.27	15.76	13.49	13.01
60/40 percent	14.99	16.08	16.60	9.15	9.21	9.15	12.72	12.28	11.69	16.83	13.71	13.15
50/50 percent	14.48	16.05	16.77	8.84	9.21	9.40	13.43	12.56	12.26	16.45	13.69	13.27

Table B.8: RMSE errors from the CNN, SVM and LSTM models for all Free State dataset combinations mixed downwards for predictions in Experiment II. The total synthetic to original data ratio is used in the proportions shown in the table. Lower RMSE percentages indicate better prediction accuracy of the model and vice-versa.

Free State Dataset Synthetic: Original (20,000:3763)	***CNN RMSE when mixed downwards (Lag1)***	***SVM RMSE when mixed downwards (Lag1)***	***LSTM RMSE when mixed downwards (Lag1)***	***CNN RMSE when mixed downwards (Lag5)***	***SVM RMSE when mixed downwards (Lag5)***	***LSTM RMSE when mixed downwards (Lag5)***	***CNN RMSE when mixed downwards (Lag14)***	***SVM RMSE when mixed downwards (Lag14)***	***LSTM RMSE when mixed downwards (Lag14)***	***CNN RMSE when mixed downwards (Lag21)***	***SVM RMSE when mixed downwards (Lag21)***	***LSTM RMSE when mixed downwards (Lag21)***
90/10 percent	15.29	15.91	15.01	9.49	9.53	9.12	13.93	12.63	12.32	15.83	14.25	13.52
80/20 percent	16.20	16.10	15.37	10.19	10.66	10.09	14.27	14.05	13.32	17.44	15.47	14.76
70/30 percent	15.54	16.18	15.46	10.99	11.65	10.96	15.09	15.07	14.57	20.09	16.76	15.35
60/40 percent	16.02	16.51	15.88	13.20	12.98	12.46	16.94	16.51	15.58	18.80	17.90	16.18
50/50 percent	17.57	17.97	17.38	15.06	14.48	13.98	18.34	17.57	17.63	21.20	19.90	17.67

Table B.9: RMSE errors from the CNN, SVM and LSTM models for all Mpumalanga dataset combinations mixed upwards for predictions in Experiment II. The total synthetic to original data ratio is used in the proportions shown in the table. Lower RMSE percentages indicate better prediction accuracy of the model and vice-versa.

Mpumalanga Dataset Synthetic: Original (20,000:3763)	***CNN RMSE when mixed upwards (Lag1)***	***SVM RMSE when mixed upwards (Lag1)***	***LSTM RMSE when mixed upwards (Lag1)***	***CNN RMSE when mixed upwards (Lag5)***	***SVM RMSE when mixed upwards (Lag5)***	***LSTM RMSE when mixed upwards (Lag5)***	***CNN RMSE when mixed upwards (Lag14)***	***SVM RMSE when mixed upwards (Lag14)***	***LSTM RMSE when mixed upwards (Lag14)***	***CNN RMSE when mixed upwards (Lag21)***	***SVM RMSE when mixed upwards (Lag21)***	***LSTM RMSE when mixed upwards (Lag21)***
90/10 percent	9.71	10.92	10.12	6.43	6.52	6.47	7.01	6.65	6.69	8.21	6.86	6.85
80/20 percent	10.00	11.06	10.43	6.49	6.57	6.51	7.30	6.75	6.68	8.21	6.91	6.80
70/30 percent	10.00	11.19	10.56	6.85	6.56	6.45	7.35	6.86	6.90	8.73	6.95	6.82
60/40 percent	9.69	11.26	10.72	6.44	6.51	6.46	7.17	6.85	7.05	8.17	7.02	7.00
50/50 percent	9.89	11.18	10.58	6.53	6.53	6.66	7.27	6.88	6.91	8.44	7.04	7.00

Table B.10: RMSE errors from the CNN, SVM and LSTM models for all Mpumalanga dataset combinations mixed downwards for predictions in Experiment II. The total synthetic to original data ratio is used in the proportions shown in the table. Lower RMSE percentages indicate better prediction accuracy of the model and vice-versa.

Mpumalanga Dataset Synthetic: Original (20,000:3763)	***CNN RMSE when mixed downwards (Lag1)***	***SVM RMSE when mixed downwards (Lag1)***	***LSTM RMSE when mixed downwards (Lag1)***	***CNN RMSE when mixed downwards (Lag5)***	***SVM RMSE when mixed downwards (Lag5)***	***LSTM RMSE when mixed downwards (Lag5)***	***CNN RMSE when mixed downwards (Lag14)***	***SVM RMSE when mixed downwards (Lag14)***	***LSTM RMSE when mixed downwards (Lag14)***	***CNN RMSE when mixed downwards (Lag21)***	***SVM RMSE when mixed downwards (Lag21)***	***LSTM RMSE when mixed downwards (Lag21)***
90/10 percent	11.46	11.18	11.37	8.11	7.65	7.65	8.10	7.45	7.65	9.45	7.74	8.01
80/20 percent	13.30	12.71	12.50	9.39	8.88	8.85	9.86	8.31	8.86	11.55	8.66	8.84
70/30 percent	14.67	13.44	13.92	11.15	10.01	10.41	10.61	8.95	9.83	12.70	9.26	9.53
60/40 percent	15.06	14.15	14.20	11.30	11.01	10.87	11.99	10.23	11.28	13.34	10.45	10.71
50/50 percent	16.56	15.75	15.92	14.36	12.77	12.66	13.32	12.09	13.02	15.22	12.07	11.93

Table B.11: RMSE errors from the CNN, SVM and LSTM models for all Northern Cape dataset combinations mixed upwards for predictions in Experiment II. The total synthetic to original data ratio is used in the proportions shown in the table. Lower RMSE percentages indicate better prediction accuracy of the model and vice-versa.

Northern Cape Dataset Synthetic: Original (20,000:3763)	***CNN RMSE when mixed upwards (Lag1)***	***SVM RMSE when mixed upwards (Lag1)***	***LSTM RMSE when mixed upwards (Lag1)***	***CNN RMSE when mixed upwards (Lag5)***	***SVM RMSE when mixed upwards (Lag5)***	***LSTM RMSE when mixed upwards (Lag5)***	***CNN RMSE when mixed upwards (Lag14)***	***SVM RMSE when mixed upwards (Lag14)***	***LSTM RMSE when mixed upwards (Lag14)***	***CNN RMSE when mixed upwards (Lag21)***	***SVM RMSE when mixed upwards (Lag21)***	***LSTM RMSE when mixed upwards (Lag21)***
90/10 percent	9.64	9.82	9.84	6.90	7.04	7.06	8.13	7.78	7.62	9.19	7.80	7.66
80/20 percent	9.62	9.74	9.71	7.05	7.07	7.08	7.91	7.74	7.32	9.54	7.71	7.29
70/30 percent	9.51	9.61	9.56	7.50	7.05	7.21	8.04	7.79	7.48	10.40	7.70	7.35
60/40 percent	9.51	9.71	9.63	7.41	7.22	7.19	8.38	7.75	7.55	9.62	7.85	7.57
50/50 percent	9.53	9.91	9.79	7.39	7.26	7.50	8.23	7.73	7.68	9.47	8.29	8.07

Table B.12: RMSE errors from the CNN, SVM and LSTM models for all Northern Cape dataset combinations mixed downwards for predictions in Experiment II. The total synthetic to original data ratio is used in the proportions shown in the table. Lower RMSE percentages indicate better prediction accuracy of the model and vice-versa.

Northern Cape Dataset Synthetic/Original (20,000/3763)	***CNN RMSE when mixed downwards (Lag1)***	***SVM RMSE when mixed downwards (Lag1)***	***LSTM RMSE when mixed downwards (Lag1)***	***CNN RMSE when mixed downwards (Lag5)***	***SVM RMSE when mixed downwards (Lag5)***	***LSTM RMSE when mixed downwards (Lag5)***	***CNN RMSE when mixed downwards (Lag14)***	***SVM RMSE when mixed downwards (Lag14)***	***LSTM RMSE when mixed downwards (Lag14)***	***CNN RMSE when mixed downwards (Lag21)***	***SVM RMSE when mixed downwards (Lag21)***	***LSTM RMSE when mixed downwards (Lag21)***
90/10 percent	9.54	9.73	9.52	6.97	7.06	6.90	8.18	7.75	7.78	10.32	7.97	7.45
80/20 percent	9.56	9.64	9.64	7.40	7.13	7.18	8.62	7.73	7.59	10.41	8.10	7.69
70/30 percent	9.72	9.70	9.73	7.61	7.24	7.21	8.56	7.73	7.63	10.01	8.33	7.86
60/40 percent	9.89	9.78	9.69	8.00	7.67	7.67	8.91	8.13	7.93	10.91	8.74	8.30
50/50 percent	10.35	10.34	10.43	8.82	8.73	8.37	9.29	8.82	8.52	10.97	9.15	8.81

Table B.13: RMSE errors from the CNN, SVM and LSTM models for all North West dataset combinations mixed upwards for predictions in Experiment II. The total synthetic to original data ratio is used in the proportions shown in the table. Lower RMSE percentages indicate better prediction accuracy of the model and vice-versa.

North West Dataset Synthetic: Original (20,000:3763)	***CNN RMSE when mixed upwards (Lag1)***	***SVM RMSE when mixed upwards (Lag1)***	***LSTM RMSE when mixed upwards (Lag1)***	***CNN RMSE when mixed upwards (Lag5)***	***SVM RMSE when mixed upwards (Lag5)***	***LSTM RMSE when mixed upwards (Lag5)***	***CNN RMSE when mixed upwards (Lag14)***	***SVM RMSE when mixed upwards (Lag14)***	***LSTM RMSE when mixed upwards (Lag14)***	***CNN RMSE when mixed upwards (Lag21)***	***SVM RMSE when mixed upwards (Lag21)***	***LSTM RMSE when mixed upwards (Lag21)***
90/10 percent	8.95	9.36	9.13	8.84	8.34	8.30	10.19	8.83	8.82	11.07	9.15	9.06
80/20 percent	8.53	9.26	8.89	8.65	8.27	8.24	9.61	8.79	8.79	10.77	9.16	9.08
70/30 percent	8.65	9.30	9.01	8.38	8.28	8.34	9.74	8.87	9.00	10.95	9.41	9.28
60/40 percent	8.90	9.32	9.01	8.62	8.18	8.18	9.71	8.87	8.90	10.88	9.35	9.20
50/50 percent	8.86	9.39	9.09	8.68	8.25	8.26	9.73	9.01	9.05	12.79	9.58	9.36

Table B.14: RMSE errors from the CNN, SVM and LSTM models for all North West dataset combinations mixed downwards for predictions in Experiment II. The total synthetic to original data ratio is used in the proportions shown in the table. Lower RMSE percentages indicate better prediction accuracy of the model and vice-versa.

North West Dataset Synthetic: Original (20,000:3763)	***CNN RMSE when mixed downwards (Lag1)***	***SVM RMSE when mixed downwards (Lag1)***	***LSTM RMSE when mixed downwards (Lag1)***	***CNN RMSE when mixed downwards (Lag5)***	***SVM RMSE when mixed downwards (Lag5)***	***LSTM RMSE when mixed downwards (Lag5)***	***CNN RMSE when mixed downwards (Lag14)***	***SVM RMSE when mixed downwards (Lag14)***	***LSTM RMSE when mixed downwards (Lag14)***	***CNN RMSE when mixed downwards (Lag21)***	***SVM RMSE when mixed downwards (Lag21)***	***LSTM RMSE when mixed downwards (Lag21)***
90/10 percent	9.86	10.10	9.73	8.66	8.30	8.25	9.81	8.84	9.03	10.99	9.29	9.27
80/20 percent	10.63	10.71	10.36	8.88	8.58	8.55	10.21	9.03	8.93	10.51	9.53	9.37
70/30 percent	11.03	11.00	10.73	9.03	8.73	8.75	10.68	9.13	9.06	10.85	9.71	9.46
60/40 percent	11.42	11.51	11.69	9.84	9.18	9.18	10.64	9.60	9.62	12.31	10.09	9.75
50/50 percent	12.45	12.04	12.01	10.44	9.73	9.81	10.79	10.00	9.84	12.33	10.52	10.09

Table B.15: RMSE errors from the CNN, SVM and LSTM models for all Gauteng dataset combinations mixed upwards for predictions in Experiment II. The total synthetic to original data ratio is used in the proportions shown in the table. Lower RMSE percentages indicate better prediction accuracy of the model and vice-versa.

Gauteng Dataset Synthetic: Original (20,000:3763)	***CNN RMSE when mixed upwards (Lag1)***	***SVM RMSE when mixed upwards (Lag1)***	***LSTM RMSE when mixed upwards (Lag1)***	***CNN RMSE when mixed upwards (Lag5)***	***SVM RMSE when mixed upwards (Lag5)***	***LSTM RMSE when mixed upwards (Lag5)***	***CNN RMSE when mixed upwards (Lag14)***	***SVM RMSE when mixed upwards (Lag14)***	***LSTM RMSE when mixed upwards (Lag14)***	***CNN RMSE when mixed upwards (Lag21)***	***SVM RMSE when mixed upwards (Lag21)***	***LSTM RMSE when mixed upwards (Lag21)***
90/10 percent	71.61	76.56	75.10	44.86	44.58	44.77	45.16	42.50	44.27	50.61	42.25	41.48
80/20 percent	70.91	76.65	75.30	46.59	44.56	44.88	45.52	42.70	42.62	49.28	42.09	40.92
70/30 percent	71.00	76.77	75.66	47.46	45.58	47.34	46.79	43.70	43.68	48.63	41.85	41.33
60/40 percent	72.05	77.19	75.58	45.93	45.57	47.06	49.09	45.41	45.60	50.20	41.69	41.99
50/50 percent	72.98	79.78	77.19	55.76	46.76	47.46	50.90	45.89	46.32	51.97	42.52	43.17

Table B16: RMSE errors from the CNN, SVM and LSTM models for all Gauteng dataset combinations mixed downwards for predictions in Experiment II. The total synthetic to original data ratio is used in the proportions shown in the table. Lower RMSE percentages indicate better prediction accuracy of the model and vice-versa.

Gauteng Dataset Synthetic: Original (20,000:3763)	***CNN RMSE when mixed downwards (Lag1)***	***SVM RMSE when mixed downwards (Lag 1)***	***LSTM RMSE when mixed downwards (Lag 1)***	***CNN RMSE when mixed downwards (Lag5)***	***SVM RMSE when mixed downwards (Lag5)***	***LSTM RMSE when mixed downwards (Lag5)***	***CNN RMSE when mixed downwards (Lag14)***	***SVM RMSE when mixed downwards (Lag14)***	***LSTM RMSE when mixed downwards (Lag14)***	***CNN RMSE when mixed downwards (Lag21)***	***SVM RMSE when mixed downwards (Lag21)***	***LSTM RMSE when mixed downwards (Lag21)***
90/10 percent	70.00	75.54	73.90	46.32	46.02	46.97	46.67	43.05	43.17.	53.62	42.43	44.45
80/20 percent	71.47	74.13	73.57	50.51	47.43	48.34	47.89	44.87	44.75	52.16	43.9	45.20
70/30 percent	72.28	73.03	74.45	49.76	48.94	49.45	52.99	46.25	46.47	54.43	45.01	47.48
60/40 percent	79.55	73.48	74.26	56.47	51.29	52.07	51.67	48.89	49.94	57.54	48.41	48.36
50/50 percent	81.24	77.09	78.01	62.69	57.26	57.12	61.18	56.80	56.82	64.34	55.48	55.42

Table B.17: RMSE errors from the CNN, SVM and LSTM models for all Eastern cape dataset combinations mixed upwards for predictions in Experiment II. The total synthetic to original data ratio is used in the proportions shown in the table. Lower RMSE percentages indicate better prediction accuracy of the model and vice-versa.

Eastern Cape Dataset ***Synthetic: Original (20,000:3763)***	***CNN RMSE when mixed upwards (Lag1)***	***SVM RMSE when mixed upwards (Lag1)***	***LSTM RMSE when mixed upwards (Lag1)***	***CNN RMSE when mixed upwards (Lag5)***	***SVM RMSE when mixed upwards (Lag5)***	***LSTM RMSE when mixed upwards (Lag5)***	***CNN RMSE when mixed upwards (Lag14)***	***SVM RMSE when mixed upwards (Lag14)***	***LSTM RMSE when mixed upwards (Lag14)***	***CNN RMSE when mixed upwards (Lag21)***	***SVM RMSE when mixed upwards (Lag21)***	***LSTM RMSE when mixed upwards (Lag21)***
90/10 percent	33.01	34.14	34.20	13.68	13.42	13.94	13.59	13.14	13.16	14.69	12.46	12.03
80/20 percent	33.87	33.43	34.01	13.62	13.46	13.62	13.89	12.96	13.01	14.96	12.56	12.07
70/30 percent	33.13	33.93	34.75	14.52	13.72	14.11	14.04	12.97	13.29	14.87	12.78	12.64
60/40 percent	32.92	33.26	34.01	13.87	13.89	14.11	14.34	13.22	13.10	15.10	12.83	12.34
50/50 percent	33.15	34.00	34.20	14.82	13.88	14.33	14.89	13.51	13.50	15.77	13.03	12.59

Table B.18: RMSE errors from the CNN, SVM and LSTM models for all Eastern cape dataset combinations mixed downwards for predictions in Experiment II. The total synthetic to original data ratio is used in the proportions shown in the table. Lower RMSE percentages indicate better prediction accuracy of the model and vice-versa.

Eastern Cape Dataset ***Synthetic: Original (20,000:3763)***	***CNN RMSE when mixed downwards (Lag1)***	***SVM RMSE when mixed downwards (Lag 1)***	***LSTM RMSE when mixed downwards (Lag 1)***	***CNN RMSE when mixed downwards (Lag5)***	***SVM RMSE when mixed downwards (Lag5)***	***LSTM RMSE when mixed downwards (Lag5)***	***CNN RMSE when mixed downwards (Lag14)***	***SVM RMSE when mixed downwards (Lag14)***	***LSTM RMSE when mixed downwards (Lag14)***	***CNN RMSE when mixed downwards (Lag21)***	***SVM RMSE when mixed downwards (Lag21)***	***LSTM RMSE when mixed downwards (Lag21)***
90/10 percent	33.74	34.58	35.04	13.79	13.53	13.57	14.16	14.80	12.93	16.13	12.89	12.67
80/20 percent	32.83	33.95	33.75	14.21	13.70	13.66	14.88	13.20	12.96	15.86	13.32	12.84
70/30 percent	32.22	33.14	33.30	14.67	14.16	14.09	15.14	13.71	14.47	17.06	13.89	13.47
60/40 percent	31.23	31.97	31.97	16.27	14.70	14.80	16.52	14.45	14.23	18.24	15.34	14.62
50/50 percent	31.49	31.41	31.25	18.57	16.83	16.77	17.54	16.76	16.55	20.22	17.53	17.27

Appendix C

RMSE for all Prediction Scenarios in Experiment III (predictions with combinations of original and synthetic data and REVAC parameter tuning)

Table C. 1: *RMSE errors from the CNN, SVM and LSTM models for all Western Cape dataset combinations augmented upwards for predictions in Experiment III. The total synthetic to original data ratio is used in the proportions shown in the table. Lower RMSE percentages indicate better prediction accuracy of the model and vice-versa.*

Western Cape Dataset Synthetic: Original (20,000:3763)	***CNN RMSE when mixed upwards (Lag1)***	***SVM RMSE when mixed upwards (Lag1)***	***LSTM RMSE when mixed upwards (Lag1)***	***CNN RMSE when mixed upwards (Lag5)***	***SVM RMSE when mixed upwards (Lag5)***	***LSTM RMSE when mixed upwards (Lag5)***	***CNN RMSE when mixed upwards (Lag14)***	***SVM RMSE when mixed upwards (Lag14)***	***LSTM RMSE when mixed upwards (Lag14)***	***CNN RMSE when mixed upwards (Lag21)***	***SVM RMSE when mixed upwards (Lag21)***	***LSTM RMSE when mixed upwards (Lag21)***
90/10 percent	69.22	76.62	69.22	46.68	48.21	46.62	56.76	57.04	54.24	59.42	59.17	58.49
80/20 percent	69.72	75.53	68.86	52.85	47.46	47.68	58.49	56.86	54.88	58.84	59.48	58.59
70/30 percent	70.23	75.78	68.94	48.64	47.49	47.03	55.65	57.42	55.62	60.96	59.02	58.78
60/40 percent	71.80	75.13	71.23	51.10	47.31	49.76	58.82	56.93	56.12	61.61	59.04	58.02
50/50 percent	72.81	82.68	73.21	49.20	47.82	48.77	56.81	56.70	57.09	65.13	60.16	59.04

Table C.2: RMSE errors from the CNN, SVM and LSTM models for all Western Cape dataset combinations augmented downwards for predictions in Experiment III. The total synthetic to original data ratio is used in the proportions shown in the table. Lower RMSE percentages indicate better prediction accuracy of the model and vice-versa.

Western Cape Synthetic: Original (20,000:3763)	***CNN RMSE when mixed downwards (Lag1)***	***SVM RMSE when mixed downwards (Lag 1)***	***LSTM RMSE when mixed downwards (Lag 1)***	***CNN RMSE when mixed downwards (Lag5)***	***SVM RMSE when mixed downwards (Lag5)***	***LSTM RMSE when mixed downwards (Lag5)***	***CNN RMSE when mixed downwards (Lag14)***	***SVM RMSE when mixed downwards (Lag14)***	***LSTM RMSE when mixed downwards (Lag14)***	***CNN RMSE when mixed downwards (Lag21)***	***SVM RMSE when mixed downwards (Lag21)***	***LSTM RMSE when mixed downwards (Lag21)***
90/10 percent	67.89	74.00	68.75	46.09	53.40	45.32	54.28	70.26	54.16	58.16	72.36	58.18
80/20 percent	70.75	72.92	68.58	46.96	58.02	46.12	57.80	78.69	56.77	58.41	82.62	57.13
70/30 percent	69.21	72.36	69.90	48.53	61.86	45.89	52.73	85.35	53.18	58.91	87.16	58.08
60/40 percent	75.10	74.31	73.94	49.98	64.75	49.36	56.44	89.79	56.63	59.84	90.47	57.71
50/50 percent	83.69	80.44	87.89	59.67	77.24	55.57	62.81	93.91	62.28	64.86	97.86	64.20

Table C.3: RMSE errors from the CNN, SVM and LSTM models for all KwaZulu Natal dataset combinations augmented upwards for predictions in Experiment III. The total synthetic to original data ratio is used in the proportions shown in the table. Lower RMSE percentages indicate better prediction accuracy of the model and vice-versa.

KwaZulu Natal Dataset Synthetic: Original (20,000:3763)	***CNN RMSE when mixed upwards (Lag1)***	***SVM RMSE when mixed upwards (Lag1)***	***LSTM RMSE when mixed upwards (Lag1)***	***CNN RMSE when mixed upwards (Lag5)***	***SVM RMSE when mixed upwards (Lag5)***	***LSTM RMSE when mixed upwards (Lag5)***	***CNN RMSE when mixed upwards (Lag14)***	***SVM RMSE when mixed upwards (Lag14)***	***LSTM RMSE when mixed upwards (Lag14)***	***CNN RMSE when mixed upwards (Lag21)***	***SVM RMSE when mixed upwards (Lag21)***	***LSTM RMSE when mixed upwards (Lag21)***
90/10 percent	43.46	47.82	43.49	20.42	20.89	20.53	23.98	19.44	20.22	25.25	21.19	20.44
80/20 percent	42.01	46.49	42.76	20.36	21.01	21.23	25.04	20.08	20.90	24.89	20.99	21.06
70/30 percent	43.62	47.41	44.66	21.38	22.36	20.96	23.18	20.25	21.70	23.49	21.09	21.97
60/40 percent	43.11	48.92	45.76	21.39	21.37	21.86	23.51	20.95	21.94	23.01	21.68	22.85
50/50 percent	44.57	50.31	46.36	21.42	21.85	22.46	24.44	22.04	22.72	23.61	22.29	23.88

Table C.4: RMSE errors from the CNN, SVM and LSTM models for all KwaZulu Natal dataset combinations augmented upwards for predictions in Experiment III. The total synthetic to original data ratio is used in the proportions shown in the table. Lower RMSE percentages indicate better prediction accuracy of the model and vice-versa.

KwaZulu Natal Dataset Synthetic: Original (20,000:3763)	***CNN RMSE when mixed downwards (Lag1)***	***SVM RMSE when mixed downwards (Lag 1)***	***LSTM RMSE when mixed downwards (Lag 1)***	***CNN RMSE when mixed downwards (Lag5)***	***SVM RMSE when mixed downwards (Lag5)***	***LSTM RMSE when mixed downwards (Lag5)***	***CNN RMSE when mixed downwards (Lag14)***	***SVM RMSE when mixed downwards (Lag14)***	***LSTM RMSE when mixed downwards (Lag14)***	***CNN RMSE when mixed downwards (Lag21)***	***SVM RMSE when mixed downwards (Lag21)***	***LSTM RMSE when mixed downwards (Lag21)***
90/10 percent	43.66	48.77	44.51	21.66	26.76	19.82	22.47	35.98	20.81	22.48	34.28	19.51
80/20 percent	43.54	48.61	43.11	22.12	31.32	22.83	24.00	45.65	20.29	21.98	42.54	22.00
70/30 percent	45.81	48.83	44.69	22.18	34.56	23.20	24.12	49.37	24.05	23.35	47.42	21.68
60/40 percent	44.18	48.79	45.16	23.51	32.48	25.11	33.87	56.76	23.71	28.89	54.38	26.11
50/50 percent	49.97	52.25	49.97	30.14	43.51	28.33	30.84	65.02	29.53	30.77	62.61	28.99

Table C.5: RMSE errors from the CNN, SVM and LSTM models for all Limpopo dataset combinations augmented upwards for predictions in Experiment III. The total synthetic to original data ratio is used in the proportions shown in the table. Lower RMSE percentages indicate better prediction accuracy of the model and vice-versa.

Limpopo Dataset Synthetic: Original (20,000:3763)	***CNN RMSE when mixed upwards (Lag1)***	***SVM RMSE when mixed upwards (Lag1)***	***LSTM RMSE when mixed upwards (Lag1)***	***CNN RMSE when mixed upwards (Lag5)***	***SVM RMSE when mixed upwards (Lag5)***	***LSTM RMSE when mixed upwards (Lag5)***	***CNN RMSE when mixed upwards (Lag14)***	***SVM RMSE when mixed upwards (Lag14)***	***LSTM RMSE when mixed upwards (Lag14)***	***CNN RMSE when mixed upwards (Lag21)***	***SVM RMSE when mixed upwards (Lag21)***	***LSTM RMSE when mixed upwards (Lag21)***
90/10 percent	4.55	4.69	4.54	3.12	2.97	3.00	3.45	3.35	3.16	3.46	3.45	3.28
80/20 percent	4.60	4.64	4.56	3.18	3.02	3.01	3.47	3.32	3.25	3.52	3.45	3.27
70/30 percent	4.57	4.69	4.56	3.27	3.10	3.02	3.45	3.39	3.27	3.46	3.51	3.24
60/40 percent	4.77	4.81	4.65	3.54	3.23	3.12	3.63	3.47	3.68	3.63	3.56	3.53
50/50 percent	4.96	4.95	4.83	3.60	3.22	3.06	3.86	3.58	3.53	3.71	3.65	3.47

Table C.6: RMSE errors from the CNN, SVM and LSTM models for all Limpopo dataset combinations augmented downwards for predictions in Experiment III. The total synthetic to original data ratio is used in the proportions shown in the table. Lower RMSE percentages indicate better prediction accuracy of the model and vice-versa.

Limpopo Dataset Synthetic: Original (20,000:3763)	***CNN RMSE when mixed downwards (Lag1)***	***SVM RMSE when mixed downwards (Lag 1)***	***LSTM RMSE when mixed downwards (Lag 1)***	***CNN RMSE when mixed downwards (Lag5)***	***SVM RMSE when mixed downwards (Lag5)***	***LSTM RMSE when mixed downwards (Lag5)***	***CNN RMSE when mixed downwards (Lag14)***	***SVM RMSE when mixed downwards (Lag14)***	***LSTM RMSE when mixed downwards (Lag14)***	***CNN RMSE when mixed downwards (Lag21)***	***SVM RMSE when mixed downwards (Lag21)***	***LSTM RMSE when mixed downwards (Lag21)***
90/10 percent	4.75	4.95	4.76	3.32	3.33	3.24	3.66	3.36	3.62	3.69	3.67	3.55
80/20 percent	4.88	5.04	4.87	3.70	3.7	3.67	3.97	3.89	3.89	3.99	3.99	3.81
70/30 percent	5.13	5.15	5.04	4.27	4.16	4.32	4.53	4.36	4.28	4.39	4.4	4.29
60/40 percent	5.78	5.82	5.69	5.13	5.02	5.22	5.23	5.17	5.20	5.35	5.5	5.42
50/50 percent	7.00	7.17	6.95	6.62	6.4	6.45	6.29	6.57	6.45	6.31	6.8	6.51

Table C.7: RMSE errors from the CNN, SVM and LSTM models for all Free State dataset combinations augmented upwards for predictions in Experiment III. The total synthetic to original data ratio is used in the proportions shown in the table. Lower RMSE percentages indicate better prediction accuracy of the model and vice-versa.

Free State Dataset Synthetic: Original (20,000:3763)	***CNN RMSE when mixed upwards (Lag1)***	***SVM RMSE when mixed upwards (Lag1)***	***LSTM RMSE when mixed upwards (Lag1)***	***CNN RMSE when mixed upwards (Lag5)***	***SVM RMSE when mixed upwards (Lag5)***	***LSTM RMSE when mixed upwards (Lag5)***	***CNN RMSE when mixed upwards (Lag14)***	***SVM RMSE when mixed upwards (Lag14)***	***LSTM RMSE when mixed upwards (Lag14)***	***CNN RMSE when mixed upwards (Lag21)***	***SVM RMSE when mixed upwards (Lag21)***	***LSTM RMSE when mixed upwards (Lag21)***
90/10 percent	14.55	14.99	14.94	8.55	8.30	8.36	11.55	12.29	11.31	12.73	14.16	13.09
80/20 percent	14.40	14.93	14.60	8.52	8.39	8.48	11.44	12.14	11.24	12.92	14.44	13.23
70/30 percent	14.51	14.75	14.44	8.60	8.34	8.41	12.04	12.43	11.14	13.51	14.56	12.92
60/40 percent	15.07	15.34	15.42	8.89	8.81	8.56	11.95	13.04	11.97	13.42	14.83	12.90
50/50 percent	14.50	15.15	15.24	8.59	8.78	8.40	11.87	13.06	11.90	13.77	14.64	13.13

Table C.8: RMSE errors from the CNN, SVM and LSTM models for all Free State dataset combinations augmented downwards for predictions in Experiment III. The total synthetic to original data ratio is used in the proportions shown in the table. Lower RMSE percentages indicate better prediction accuracy of the model and vice-versa.

Free State Dataset Synthetic: Original (20,000:3763)	***CNN RMSE when mixed downwards (Lag1)***	***SVM RMSE when mixed downwards (Lag 1)***	***LSTM RMSE when mixed downwards (Lag 1)***	***CNN RMSE when mixed downwards (Lag5)***	***SVM RMSE when mixed downwards (Lag5)***	***LSTM RMSE when mixed downwards (Lag5)***	***CNN RMSE when mixed downwards (Lag14)***	***SVM RMSE when mixed downwards (Lag14)***	***LSTM RMSE when mixed downwards (Lag14)***	***CNN RMSE when mixed downwards (Lag21)***	***SVM RMSE when mixed downwards (Lag21)***	***LSTM RMSE when mixed downwards (Lag21)***
90/10 percent	14.89	16.18	14.76	9.24	9.83	8.99	12.51	12.70	11.87	13.83	14.22	13.47
80/20 percent	16.41	16.33	15.29	9.84	10.75	9.92	14.47	13.91	13.22	15.71	15.24	15.53
70/30 percent	15.31	16.38	15.77	10.65	11.63	11.00	15.34	14.71	14.59	17.50	16.27	15.50
60/40 percent	16.05	16.65	15.87	12.14	12.88	12.22	16.04	15.76	15.99	19.32	17.37	17.10
50/50 percent	17.31	16.18	17.27	13.33	14.16	13.94	17.63	16.79	17.76	18.94	18.72	18.10

Table C.9: RMSE errors from the CNN, SVM and LSTM models for all Mpumalanga dataset combinations augmented upwards for predictions in Experiment III. The total synthetic to original data ratio is used in the proportions shown in the table. Lower RMSE percentages indicate better prediction accuracy of the model and vice-versa.

Mpumalanga Dataset Synthetic: Original (20,000:3763)	***CNN RMSE when mixed upwards (Lag1)***	***SVM RMSE when mixed upwards (Lag1)***	***LSTM RMSE when mixed upwards (Lag1)***	***CNN RMSE when mixed upwards (Lag5)***	***SVM RMSE when mixed upwards (Lag5)***	***LSTM RMSE when mixed upwards (Lag5)***	***CNN RMSE when mixed upwards (Lag14)***	***SVM RMSE when mixed upwards (Lag14)***	***LSTM RMSE when mixed upwards (Lag14)***	***CNN RMSE when mixed upwards (Lag21)***	***SVM RMSE when mixed upwards (Lag21)***	***LSTM RMSE when mixed upwards (Lag21)***
90/10 percent	9.66	9.93	10.12	6.50	6.18	6.39	6.80	7.05	6.49	6.72	7.41	6.88
80/20 percent	9.72	10.03	10.09	6.52	6.22	6.46	6.92	7.10	6.69	6.92	7.49	7.01
70/30 percent	10.16	10.14	10.42	6.70	6.24	6.69	6.91	7.21	6.94	6.88	7.53	6.82
60/40 percent	9.71	10.20	10.28	6.44	6.10	6.70	6.77	7.26	6.70	6.99	7.59	6.84
50/50 percent	9.76	10.22	10.01	6.32	6.20	6.68	6.88	7.18	6.68	7.31	7.63	6.77

Table C.10: RMSE errors from the CNN, SVM and LSTM models for all Mpumalanga dataset combinations augmented downwards for predictions in Experiment III. The total synthetic to original data ratio is used in the proportions shown in the table. Lower RMSE percentages indicate better prediction accuracy of the model and vice-versa.

Mpumalanga Dataset Synthetic: Original (20,000:3763)	***CNN RMSE when mixed downwards (Lag1)***	***SVM RMSE when mixed downwards (Lag 1)***	***LSTM RMSE when mixed downwards (Lag 1)***	***CNN RMSE when mixed downwards (Lag5)***	***SVM RMSE when mixed downwards (Lag5)***	***LSTM RMSE when mixed downwards (Lag5)***	***CNN RMSE when mixed downwards (Lag14)***	***SVM RMSE when mixed downwards (Lag14)***	***LSTM RMSE when mixed downwards (Lag14)***	***CNN RMSE when mixed downwards (Lag21)***	***SVM RMSE when mixed downwards (Lag21)***	***LSTM RMSE when mixed downwards (Lag21)***
90/10 percent	11.53	11.71	11.47	7.64	7.66	7.64	7.73	7.56	7.77	8.25	7.85	8.07
80/20 percent	13.32	12.66	12.43	8.60	9.01	9.04	8.86	8.56	8.59	9.85	8.94	9.28
70/30 percent	14.19	13.45	13.70	10.10	10.31	10.19	9.53	9.21	9.32	10.03	9.63	9.47
60/40 percent	14.92	14.23	13.88	10.56	11.27	10.45	11.06	10.53	10.90	10.90	10.86	10.46
50/50 percent	15.89	16.01	15.64	12.38	13.12	12.03	13.07	12.42	13.25	13.40	12.39	12.02

Table C.11: RMSE errors from the CNN, SVM and LSTM models for all Northern Cape dataset combinations augmented upwards for predictions in Experiment III. The total synthetic to original data ratio is used in the proportions shown in the table. Lower RMSE percentages indicate better prediction accuracy of the model and vice-versa.

Northern Cape Dataset ***Synthetic: Original (20,000:3763)***	***CNN RMSE when mixed upwards (Lag1)***	***SVM RMSE when mixed upwards (Lag1)***	***LSTM RMSE when mixed upwards (Lag1)***	***CNN RMSE when mixed upwards (Lag5)***	***SVM RMSE when mixed upwards (Lag5)***	***LSTM RMSE when mixed upwards (Lag5)***	***CNN RMSE when mixed upwards (Lag14)***	***SVM RMSE when mixed upwards (Lag14)***	***LSTM RMSE when mixed upwards (Lag14)***	***CNN RMSE when mixed upwards (Lag21)***	***SVM RMSE when mixed upwards (Lag21)***	***LSTM RMSE when mixed upwards (Lag21)***
90/10 percent	9.38	9.64	9.60	6.95	6.83	6.88	7.58	7.84	7.42	7.55	8.03	7.39
80/20 percent	9.41	9.58	9.60	7.00	6.87	7.03	7.44	7.79	7.31	7.71	7.93	7.24
70/30 percent	9.55	9.45	9.60	7.00	6.85	6.98	7.61	7.80	7.33	7.65	7.92	7.60
60/40 percent	9.40	9.54	9.56	7.16	7.02	7.20	7.49	7.81	7.38	7.58	8.15	7.45
50/50 percent	9.46	9.68	9.68	7.12	7.06	7.30	7.59	7.83	7.80	7.94	8.58	7.82

Table C.12: RMSE errors from the CNN, SVM and LSTM models for all Northern Cape dataset combinations augmented downwards for predictions in Experiment III. The total synthetic to original data ratio is used in the proportions shown in the table. Lower RMSE percentages indicate better prediction accuracy of the model and vice-versa.

Northern Cape Dataset ***Synthetic: Original (20,000:3763)***	***CNN RMSE when mixed downwards (Lag1)***	***SVM RMSE when mixed downwards (Lag 1)***	***LSTM RMSE when mixed downwards (Lag 1)***	***CNN RMSE when mixed downwards (Lag5)***	***SVM RMSE when mixed downwards (Lag5)***	***LSTM RMSE when mixed downwards (Lag5)***	***CNN RMSE when mixed downwards (Lag14)***	***SVM RMSE when mixed downwards (Lag14)***	***LSTM RMSE when mixed downwards (Lag14)***	***CNN RMSE when mixed downwards (Lag21)***	***SVM RMSE when mixed downwards (Lag21)***	***LSTM RMSE when mixed downwards (Lag21)***
90/10 percent	9.51	9.64	9.57	7.00	6.98	6.94	7.70	7.88	7.50	8.10	8.23	7.61
80/20 percent	9.68	9.56	9.61	7.02	7.09	7.15	7.87	7.86	7.44	8.09	8.36	7.66
70/30 percent	9.71	9.64	9.66	7.28	7.20	7.25	7.73	7.87	7.62	8.21	8.59	7.99
60/40 percent	9.66	9.72	9.75	7.71	7.64	7.66	8.26	8.26	7.82	8.57	8.99	8.26
50/50 percent	10.32	10.28	10.37	8.42	8.32	8.23	8.96	8.88	8.36	9.50	9.48	8.67

Table C.13: RMSE errors from the CNN, SVM and LSTM models for all North West dataset combinations augmented upwards for predictions in Experiment III. The total synthetic to original data ratio is used in the proportions shown in the table. Lower RMSE percentages indicate better prediction accuracy of the model and vice-versa.

North West Dataset Synthetic: Original (20,000:3763)	***CNN RMSE when mixed upwards (Lag1)***	***SVM RMSE when mixed upwards (Lag1)***	***LSTM RMSE when mixed upwards (Lag1)***	***CNN RMSE when mixed upwards (Lag5)***	***SVM RMSE when mixed upwards (Lag5)***	***LSTM RMSE when mixed upwards (Lag5)***	***CNN RMSE when mixed upwards (Lag14)***	***SVM RMSE when mixed upwards (Lag14)***	***LSTM RMSE when mixed upwards (Lag14)***	***CNN RMSE when mixed upwards (Lag21)***	***SVM RMSE when mixed upwards (Lag21)***	***LSTM RMSE when mixed upwards (Lag21)***
90/10 percent	8.71	9.07	8.86	8.56	8.15	8.21	9.02	8.91	8.83	9.86	9.53	9.47
80/20 percent	8.59	8.95	9.01	8.68	8.09	8.01	8.88	8.90	8.93	9.42	9.57	9.27
70/30 percent	8.77	9.00	8.79	8.29	8.07	8.08	9.31	8.95	8.94	9.83	9.80	9.46
60/40 percent	8.76	9.05	8.91	8.08	8.01	8.02	8.86	8.96	8.79	9.59	9.68	9.06
50/50 percent	8.69	9.10	8.99	8.16	8.11	8.22	9.04	9.12	8.90	9.62	9.86	9.39

Table C.14: RMSE errors from the CNN, SVM and LSTM models for all North West dataset combinations augmented downwards for predictions in Experiment III. The total synthetic to original data ratio is used in the proportions shown in the table. Lower RMSE percentages indicate better prediction accuracy of the model and vice-versa.

North West Dataset Synthetic: Original (20,000:3763)	***CNN RMSE when mixed downwards (Lag1)***	***SVM RMSE when mixed downwards (Lag 1)***	***LSTM RMSE when mixed downwards (Lag 1)***	***CNN RMSE when mixed downwards (Lag5)***	***SVM RMSE when mixed downwards (Lag5)***	***LSTM RMSE when mixed downwards (Lag5)***	***CNN RMSE when mixed downwards (Lag14)***	***SVM RMSE when mixed downwards (Lag14)***	***LSTM RMSE when mixed downwards (Lag14)***	***CNN RMSE when mixed downwards (Lag21)***	***SVM RMSE when mixed downwards (Lag21)***	***LSTM RMSE when mixed downwards (Lag21)***
90/10 percent	9.44	9.71	9.65	8.30	8.14	8.21	9.02	9.01	8.88	9.57	9.75	9.17
80/20 percent	10.18	10.29	10.52	8.64	8.44	8.55	9.08	9.20	8.82	9.88	9.95	9.30
70/30 percent	10.60	10.60	10.75	8.79	8.63	8.61	9.27	9.37	8.93	10.13	10.17	9.36
60/40 percent	11.38	11.18	11.60	9.29	9.16	9.15	9.81	9.88	9.38	10.50	10.53	9.79
50/50 percent	12.05	11.72	12.42	9.81	9.71	9.71	10.35	10.30	9.93	11.41	11.01	10.28

Table C.15: RMSE errors from the CNN, SVM and LSTM models for all Gauteng dataset combinations augmented upwards for predictions in Experiment III. The total synthetic to original data ratio is used in the proportions shown in the table. Lower RMSE percentages indicate better prediction accuracy of the model and vice-versa.

Gauteng Dataset Synthetic: Original (20,000:3763)	***CNN RMSE when mixed upwards (Lag1)***	***SVM RMSE when mixed upwards (Lag1)***	***LSTM RMSE when mixed upwards (Lag1)***	***CNN RMSE when mixed upwards (Lag5)***	***SVM RMSE when mixed upwards (Lag5)***	***LSTM RMSE when mixed upwards (Lag5)***	***CNN RMSE when mixed upwards (Lag14)***	***SVM RMSE when mixed upwards (Lag14)***	***LSTM RMSE when mixed upwards (Lag14)***	***CNN RMSE when mixed upwards (Lag21)***	***SVM RMSE when mixed upwards (Lag21)***	***LSTM RMSE when mixed upwards (Lag21)***
90/10 percent	73.17	75.68	71.56	45.98	43.89	44.51	43.29	42.39	42.94	43.17	42.39	41.76
80/20 percent	73.47	75.75	71.23	45.60	43.91	44.72	46.60	42.69	43.99	43.86	42.31	40.78
70/30 percent	74.25	75.81	72.63	47.05	44.81	45.78	45.12	43.78	44.01	43.26	42.19	42.68
60/40 percent	73.85	76.34	72.75	46.50	44.83	44.84	47.02	45.45	45.44	43.68	42.04	41.25
50/50 percent	74.99	78.83	73.07	47.42	46.08	45.46	48.27	45.90	44.06	45.31	42.90	43.15

Table C.16: RMSE errors from the CNN, SVM and LSTM models for all Gauteng dataset combinations augmented downwards for predictions in Experiment III. The total synthetic to original data ratio is used in the proportions shown in the table. Lower RMSE percentages indicate better prediction accuracy of the model and vice-versa.

Gauteng Dataset Synthetic: Original (20,000:3763)	***CNN RMSE when mixed downwards (Lag1)***	***SVM RMSE when mixed downwards (Lag 1)***	***LSTM RMSE when mixed downwards (Lag 1)***	***CNN RMSE when mixed downwards (Lag5)***	***SVM RMSE when mixed downwards (Lag5)***	***LSTM RMSE when mixed downwards (Lag5)***	***CNN RMSE when mixed downwards (Lag14)***	***SVM RMSE when mixed downwards (Lag14)***	***LSTM RMSE when mixed downwards (Lag14)***	***CNN RMSE when mixed downwards (Lag21)***	***SVM RMSE when mixed downwards (Lag21)***	***LSTM RMSE when mixed downwards (Lag21)***
90/10 percent	70.92	74.66	75.12	47.11	45.47	47.46	44.13	43.03	45.27	43.70	42.75	44.12
80/20 percent	71.04	73.49	74.01	48.17	47.03	48.69	45.85	44.49	46.39	50.60	44.31	45.68
70/30 percent	71.65	72.47	73.63	51.49	48.96	49.49	49.36	46.32	47.12	47.59	45.45	46.43
60/40 percent	76.25	73.09	73.56	52.26	51.45	51.29	50.27	49.25	49.06	50.56	48.77	48.84
50/50 percent	77.83	77.01	78.41	59.22	57.75	56.36	59.00	57.06	55.72	57.09	55.82	56.01

Table C.17: RMSE errors from the CNN, SVM and LSTM models for all Eastern Cape dataset combinations augmented upwards for predictions in Experiment III. The total synthetic to original data ratio is used in the proportions shown in the table. Lower RMSE percentages indicate better prediction accuracy of the model and vice-versa.

Eastern Cape Dataset Synthetic: Original (20,000:3763)	***CNN RMSE when mixed upwards (Lag1)***	***SVM RMSE when mixed upwards (Lag1)***	***LSTM RMSE when mixed upwards (Lag1)***	***CNN RMSE when mixed upwards (Lag5)***	***SVM RMSE when mixed upwards (Lag5)***	***LSTM RMSE when mixed upwards (Lag5)***	***CNN RMSE when mixed upwards (Lag14)***	***SVM RMSE when mixed upwards (Lag14)***	***LSTM RMSE when mixed upwards (Lag14)***	***CNN RMSE when mixed upwards (Lag21)***	***SVM RMSE when mixed upwards (Lag21)***	***LSTM RMSE when mixed upwards (Lag21)***
90/10 percent	32.94	34.01	32.97	13.20	13.28	13.08	12.82	13.05	13.65	12.62	12.61	12.64
80/20 percent	32.63	33.26	32.73	13.46	13.30	13.29	13.69	12.98	13.93	12.18	12.72	12.12
70/30 percent	33.02	33.78	33.23	14.15	13.52	13.86	13.05	13.04	13.25	12.27	12.90	12.32
60/40 percent	32.58	33.10	32.97	13.64	13.71	13.61	13.31	13.24	13.31	13.21	12.90	12.65
50/50 percent	33.06	33.78	33.55	13.76	13.74	13.72	13.71	13.60	13.78	13.29	13.21	13.31

Table C.18: RMSE errors from the CNN, SVM and LSTM models for all Eastern Cape dataset combinations augmented downwards for predictions in Experiment III. The total synthetic to original data ratio is used in the proportions shown in the table. Lower RMSE percentages indicate better prediction accuracy of the model and vice-versa.

Eastern Cape Dataset Synthetic: Original (20,000:3763)	***CNN RMSE when mixed downwards (Lag1)***	***SVM RMSE when mixed downwards (Lag 1)***	***LSTM RMSE when mixed downwards (Lag 1)***	***CNN RMSE when mixed downwards (Lag5)***	***SVM RMSE when mixed downwards (Lag5)***	***LSTM RMSE when mixed downwards (Lag5)***	***CNN RMSE when mixed downwards (Lag14)***	***SVM RMSE when mixed downwards (Lag14)***	***LSTM RMSE when mixed downwards (Lag14)***	***CNN RMSE when mixed downwards (Lag21)***	***SVM RMSE when mixed downwards (Lag21)***	***LSTM RMSE when mixed downwards (Lag21)***
90/10 percent	33.39	34.41	33.83	13.51	13.40	13.86	12.99	13.16	13.10	12.73	13.02	12.51
80/20 percent	32.83	33.77	33.20	14.06	13.58	13.76	13.02	13.25	13.37	13.08	13.50	12.94
70/30 percent	31.66	32.98	32.36	14.21	14.08	14.23	13.58	13.78	13.88	13.65	14.06	13.22
60/40 percent	30.33	31.81	31.23	14.75	14.66	14.67	14.71	14.65	14.59	14.99	15.53	14.96
50/50 percent	30.56	31.34	30.88	16.78	16.85	16.90	16.92	17.03	16.48	17.59	17.97	17.53

Appendix D

Final Parameters for the Grid Search Tuner

Table D 1: *Final parameters for the CNN, SVM and LSTM models for each Province with Grid the search tuning.*

Parameters	Western Cape	Eastern Cape	Gauteng	Northern Cape	North West	Mpumalanga	KwaZulu Natal	Free State	Limpopo
C (SVM)	29.747	2.62	8.464	1.333	7.392	10.234	90.671	2.322	38.095
Gamma (SVM)	0.005	0.004	0.009	0.011	0.005	0.002	0.002	0.001	0.001
Neurons (LSTM)	(16,100,100)	(12)	(6,18,32)	(16)	(50,18,32)	(12,100)	(28,100,12)	(16,18)	(64)
No of epochs (LSTM)	120	100	120	40	100	40	50	120	50
Batch size (LSTM)	32	32	4	16	64	32	4	32	18
No of stacked LSTM layers	3	1	1	1	3	2	3	2	1
Learning rate (LSTM)	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001
Convolution layers (CNN)	2	3	2	1	3	3	3	3	2
Kernel size (CNN)	(32,32)	(18,6,6)	(28,6)	(12)	(24,16,18)	(28,32,18)	(24,16,16)	(64,6,12)	(18,64)
No of epochs (CNN)	100	50	50	40	40	50	60	50	40
Pool size (CNN)	1	1	1	1	1	1	1	1	1
Batch size (CNN)	16	64	32	18	4	16	64	32	64
Learning rate (CNN)	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001

Table D2: Final parameters for the CNN, SVM and LSTM models for each Province with the REVAC search tuning.

Parameters	**Western Cape**	**Eastern Cape**	**Gauteng**	**Northern Cape**	**North West**	**Mpumalanga**	**KwaZulu Natal**	**Free State**	**Limpopo**
C (SVM)	17.847	0.566	2.894	0.697	5.907	0.049	41.524	3.276	10.331
Gamma (SVM)	0.002	0.004	0.004	0.008	0.005	0.005	0.001	0.001	0.003
Neurons (LSTM)	(6)	(12,6,32)	(12)	(12)	(16)	(50,18)	(16,64,12)	(28,12)	(24,32)
No of epochs (LSTM)	40	50	60	100	100	40	70	50	40
Batch size (LSTM)	64	18	64	18	32	64	18	64	18
No of stacked LSTM layers	1	3	1	1	1	1	3	2	2
Learning rate (LSTM)	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001
Convolution layers (CNN)	2	2	1	2	2	3	3	3	1
Kernel size (CNN)	(12,16)	(12,12)	(6)	(12,16)	(32,32)	(16,18,16)	(6,28,24)	(24,6,28)	(12)
No of epochs (CNN)	100	100	60	150	50	40	150	50	50
Pool size (CNN)	1	1	1	1	1	1	1	1	1
Batch size (CNN)	64	18	16	16	64	64	32	16	4
Learning rate (CNN)	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001