Assessment Cover Sheet

A signed and completed cover sheet must be electronically attached to all work submitted for assessment.

Work submitted without a cover sheet will not be marked.

Student Name:	James Mulhall
Student Number:	40296869
Word Count:	15466

Declaration of academic integrity

I declare that I have read the Queen's University regulations on plagiarism, and that the attached submission is my own original work.

It is not similar in content to, nor based on the work of others, whether published or unpublished, except with full and proper acknowledgement.

I consent to my work being submitted to the plagiarism software used by Queen's University Belfast.

Signed: James Mulhall

Date: 29/09/2021

MACHINE-LEARNING MODELS FOR THE PREDICTION AND ANALYSIS OF CLIMATE-SENSITIVE DISEASES IN VIETNAM

James Mulhall

BSc Pharmacology (Hons)

Submitted in partial fulfilment of the requirements for the degree of:

MSc Bioinformatics and Computational Genomics

School of Medicine, Dentistry, and Biomedical Sciences

Queen's University Belfast

Primary supervisor: Dr. Son T. Mai

29/09/2021

Word Count: 15466

Preface

This dissertation has been submitted in partial fulfilment of the requirements for the MSc Bioinformatics and Computational Genomics degree at Queen's University Belfast. I confirm that this work is my own and has not been submitted for any other qualification. Some of the content regarding diarrhoea forecasting is intended to be rewritten for publication at a later data, and some of the work pertaining dengue fever is under review for publication in the article titled below:

Van-Hau Nguyen, Tran Thi Tuyet Hanh, Van-Chien Nguyen, James Mulhall, Hoang Van Minh, Trung Q. Duong, Nguyen Thi Trang Nhung, Vu Hoang Lan, Hoang Ba Minh, Do Cuong, Nguyen Ngoc Bich, Nguyen Huu Quyen, Tran Nu Quy Linh, Nguyen Thi Tho, Ngu Duy Nghia, Le Van Quoc Anh, Diep Phan, Quoc Viet Hung Nguyen, Son T. Mai. Deep learning models for forecasting dengue fever based on climate data in Vietnam. Under Review.

Acknowledgements

I would like to first express my sincere gratitude to my supervisor, Dr. Thai Son Mai, for his guidance and mentoring throughout this project. Son's (virtual) door has been open at all times, and he has been especially accommodating as I have developed my machine learning and epidemiological skills throughout the past many months.

I would also like to thank the rest of our research team in Vietnam, with special note to Mr. Van-Chien Nguyen from Hanoi University of Science and Technology, Dr. Van-Hau Nguyen from Hungyen University of Technology and Education, and Associate Professor Tran Thi Tuyet-Hanh from Hanoi University of Public Health. Chien and Hau's machine learning expertise has been invaluable to this project, and Hanh's role in the project design and path to publication has been indispensable.

Finally, I would like to thank my friends and family for their support during the project, and their patience while listening to my many ramblings about the minute details of infectious disease epidemiology and prediction models.

Abstract

Introduction: Dengue fever and diarrhoeal disease pose significant threats to national morbidity and mortality in Vietnam. Previous works have identified significant associations between meteorological factors and these infectious diseases, and developed climate-based prediction models. However, the selection of accurate long term prediction models in Vietnam is limited. Here, we developed climate factor-based traditional machine learning and deep learning models, and assessed them on forecasting dengue fever and diarrhoea incidence and outbreaks one to three months in advance.

Methods: Convolutional Neural Network (CNN), Long Short-Term Memory (LSTM), attention mechanism-enhanced LSTM (LSTM-ATT), and Transformer models were compared with a selection of five traditional machine learning models on dengue fever forecasting. Models used a climate dataset of 12 meteorological factors covering temperature, rainfall, humidity, evaporation, and sunshine hours as predictors, and were evaluated in 20 Vietnamese provinces. In the second part of the project addressing diarrhoea forecasting, a Tree-structured Parzen Estimator from Optuna was used for improved hyperparameter optimisation for the highest performing deep learning models, as well as for univariate and multivariate Seasonal Autoregressive Integrated Moving Average (SARIMA) models. These were evaluated on the five provinces with the highest diarrhoea rates as a proxy for poor clean water and sanitation infrastructure.

Results: Overall, LSTM-ATT had the lowest errors for dengue fever predictions, with average place-rankings of 1.60 and 1.95 for root mean square error (RMSE) and mean absolute error (MAE) based scoring. Mean absolute percentage errors (MAPEs) of 38.4% and above were reported. Errors increased for most, but not all provinces when forecasting multiple months ahead. LSTM-ATT also displayed the lowest MAE-based ranking for diarrhoea forecasting (2.2), with MAPEs of as low as 8.43% for one-month ahead predictions and 10.8% for three-month ahead predictions. CNN models displayed strong performance, too, with the lowest RMSE-based ranking of 2.2. SARIMA models were generally worse, but occasionally outperformed deep learning models, and identified significant lagged associations between diarrhoea rates and influenza rates, minimum absolute temperature, total rainfall, and sunshine hours.

Conclusions: After analysing a selection of climate factor-based machine learning models, LSTM-ATT displayed the highest performance in forecasting dengue fever, and it performed similarly to CNNs in forecasting diarrhoea rates. While dengue fever models had high MAPEs, these were inflated due to the presence of observed rates of zero for many months. To the best of our knowledge, this is the first study to present long-term climate-based diarrhoea forecasting models, and the first example of prediction models for long-term diarrhoea forecasting in Vietnam. Overall, deep learning models show strong potential for the development of early-warning systems for infectious disease outbreaks in Vietnam.

Table of Contents

1. Introduction

1.1 Vietnam: Geography, Climate, and Disease Susceptibility

Vietnam is a climatically diverse country due to large national variations in latitude, altitude, and susceptibility to coastal weather and flooding effects. As such, susceptibility to infectious diseases and the specific effects of weather on incidence rates may vary between provinces. The Southeast Asian country stretches across 15 degrees of latitude, with 3,260km of mainland coastline and thousands of islands (Boateng, 2012). It shares borders with the East Sea, China, Laos, and Cambodia. Vietnam is separated into three main regions—northern, central, and southern Vietnam—with further divisions into seven subregions: the Northwest, the Northeast, the Red River Delta, the North Central Coast, the South Central Coast, the Central Highlands, the Southeast, and the Mekong River Delta. Throughout these regions are 63 provinces, made up of 58 true provinces and 5 municipalities which hold equal administrative level (International Organization for Standardization, 2020). Over three quarters of the country is made up of hilly terrain from 100–1000m, with mountains in the North reaching 3,143m in height (Socialist Republic of Viet Nam, 2010). However, over 70% is at an altitude under 500m. A quarter is made up of plains, primarily the Red River Delta in the north and the Mekong river delta in the South. With regards to weather patterns, the North has a subtropical climate with four standard seasons, whereas the South has a tropical climate with a wet season and a dry season. The entire country experiences high rainfall, temperatures, and humidity. However, the Southern provinces are hotter, rainfall is higher in the mountainous provinces (e.g., the Northwest, Northeast, and Central Highlands), and the central coastal provinces experience frequent hurricanes and storms (FAO, 2011).

Vietnam is significantly vulnerable to extreme weather events, climate change, and climate-sensitive infectious diseases. Over 74% of people in Vietnam are impacted by climate vulnerability. The poor and those living near coastal regions such as the Mekong River Delta are particularly vulnerable (UNICEF, 2018). The high risk associated with islands and coastal regions is partially due to rising sea levels, which have elevated by over 20cm since the late 1950s (Institute of Strategy and Policy on Natural Resources and Environment, 2009). Moreover, extreme weather events can damage healthcare infrastructure and limit the availability of treatment for infectious diseases in Vietnam. Heavy rainfall and floods can wash contaminants into drinking water, leading to increased diarrhoea rates, or create breeding grounds for mosquitoes, causing dengue fever (DF) epidemics. At the other end of the spectrum, droughts can cause concentration of pathogens in water sources leading to increased cases of diarrhoeal diseases. For the Intergovernmental Panel on Climate Change's Representative Concentration Pathway 4.5 (RCP4.5), more frequent severe typhoons and droughts, longer monsoon seasons, and a sea level rise of 55cm are projected by the end of the 21st century in Vietnam. Temperatures are forecast to rise by approximately 2.2°C in Northern regions and 1.8°C in Southern regions, and annual rainfall by 5–15mm (Tuyet-Hanh et al., 2018a). Additionally, longer summers, and winters shortened by 1–2 months, may be commonplace in the North (Socialist Republic of Viet Nam, 2010). These climate adaptations are projected to worsen the impact of DF, diarrhoea, and other communicable diseases in Vietnam (Tuyet-Hanh et al., 2018a).

1.2 Dengue Fever

1.2.1 Aetiology & Burden

DF is a mosquito-borne neglected tropical disease caused by infection with the dengue virus. The main vectors which spread DF are *Aedes* mosquitos—*Ae. aegypti* and, to a lesser extent, *Ae. albopictus* (Higa et al., 2010; Wilke et al., 2019). The dengue virus can be passed on from an infected person if a mosquito feeds on them, after which the virus replicates and spreads to tissues throughout the mosquito including the salivary glands through which transmission is possible (World Health Organisation, 2020). At 25–28°C, this extrinsic incubation period is between 8 and 12 days, though this is temperature dependent (Tjaden et al., 2013). Mosquitos remain virulent for life, and the virus can be transmitted through saliva when an infected mosquito bites a new human host. The World Health Organisation (2020) categorises infections as either dengue or severe dengue, though the former can progress to the latter. Dengue can be asymptomatic or present with severe flu-like symptoms including fever, severe headaches, body pains, nausea, and vomiting after an intrinsic incubation period of 4 to 10 days. Severe dengue is characterised by respiratory distress, organ failure, internal bleeding, and risk of death. There are currently no specific treatments for DF. One vaccine, Dengvaxia®, was approved in 2015 but comes with severe limitations and therefore limited utility; it is only recommended in use for persons 9–45 years old with a confirmed past case of DF (European Medicines Agency, 2020).

DF represents a significant growing global health burden, with an estimated 390 million cases per year—294 million of which are asymptomatic. Asia is affected the most with ~70% of infections, while the Americas also represent a high risk zone (Bhatt et al., 2013). On average, 81 thousand cases were reported annually in Vietnam between 1997 and 2016, indicating a

substantial public health risk. Climate change is predicted to lead to a more severe situation in Vietnam, due to temperatures throughout the country and particularly in the central regions warming to ranges favourable for DF transmission (Tuyet-Hanh et al., 2018a).

1.2.2 The Dengue Virus

The dengue virus is a single-stranded positive-sense RNA Flavivirus (Aguas et al., 2019). Four serotypes of the dengue virus are commonly referred to in the literature (Bharaj et al., 2008; Villabona-Arenas et al., 2014), with a putative fifth described as early as 2013 (Mustafa et al., 2015; Normile, 2013). Dengue serotypes co-circulate in highly affected regions such as Vietnam, with the dominant serotype changing over time. This is proposed to contribute to the seasonal and multi-annual variability in infection dynamics. Infection with one serotype can lead to either cross-protection, where an individual gains immunity to other serotypes, or cross-enhancement, where the individual is more likely to suffer from severe dengue and increased infectivity after heterologous infection (ten Bosch et al., 2016). Cross-enhancement is believed to be caused by antibody-dependent enhancement (ADE), a process by which secondary dengue infection leads to downregulation of antiviral processes and facilitation of viral uptake into cells (Halstead, 2014). The reality of secondary heterologous dengue infection is highly complex, with the degree of protective or enhancing effects dependent on the specific serotypes involved (Aguas et al., 2019).

1.2.3 Associations between Climate Factors and Dengue Fever

Past works have investigated the relationships between DF and climate factors such as temperature and rainfall; an understanding of such research has utility in designing and interpreting effective forecasting models. Positive correlations between minimum temperature and DF have been found for 1–2 month lags (Colón-González et al., 2011; Lowe et al., 2018; Phung et al., 2015c; Wang et al., 2014), with negative correlations for the same month also being reported (Wang et al., 2014). At 0–2.5 month lags, positive correlations were also found with average temperature in most (Do et al., 2014; Lee et al., 2017b; Pham et al., 2011; Phuong et al., 2016) but not all (Tuyet-Hanh et al., 2018b) studies reviewed. Regarding rainfall, positive associations with DF are commonly reported for lag-times of 0–3 months (Do et al., 2014; Lee et al., 2017b; Lowe et al., 2018; Pham et al., 2011; Phung et al., 2015b; Phuong et al., 2016), though negative associations (Wang et al., 2014) and lack of association (Colón-González et al., 2011; Tuyet-Hanh et al., 2018b) are also reported. Therefore, temperature and rainfall appear to be strong candidates for predictors in DF forecasting models, in spite of the mixed results.

While rainfall and temperature have been the focus of most analyses, humidity, evaporation, sunshine hours, wind speed, and El Niño events may also affect DF incidence. Relative and minimum humidity have been shown to be associated with DF for lags of 1–3 months (Phung et al., 2015b; Wang et al., 2014). Similarly, evaporation has been shown to be associated with DF during the same month (Tuyet-Hanh et al., 2018b). The impact of sunshine hours is unclear, with works in support of both positive (Tuyet-Hanh et al., 2018b) and negative correlations with DF (Pham et al., 2011). Lastly, wind speed (Wang et al., 2014) and El Niño events (Colón-

González et al., 2011) have been shown to have inverse correlations with rates of DF in the same month.

Risk mapping has highlighted provinces in Southern Vietnam as particularly high-risk for dengue incidence, while in the North, Hà Nội has been shown to experience notably higher incidence than proximal provinces (Bett et al., 2019). As such, many of the studies in Vietnam elucidating the effect of environmental variables on DF have focused on Hà Nội and the Southern provinces. The variation in relationships between climate and DF in Vietnam and further afield therefore may partially be explained by the difference in study locations

1.2.4 Dengue Fever Prediction Models

Statistical and machine learning models have previously been applied to the prediction of DF from climate factors, a selection of which are described here. In Cần Thơ, Vietnam, Phung et al. (2015b) compared DF prediction models, using temperature, humidity, and rainfall as climate variables. Multiple regression, Seasonal Autoregressive Integrated Moving Average (SARIMA), and Poisson distributed lag models were evaluated using mean absolute percentage error (MAPE), identifying the Poisson distributed lag model as the best overall predictor of disease with a MAPE of 9%. Bett and colleagues (2019) used Vietnamese land cover and altitude data in addition to climate data to develop a DF prediction model based on hierarchical spatial Bayesian statistics using integrated nested Laplace approximation. A Besag-York-Mollie conditional autoregressive model was included, which accounted for the spatial effects of neighbouring provinces on case numbers. This led to the identification of minimum

temperature, rainfall, urban land cover, altitude, spatial conditional autoregression, and temporal autocorrelation as significant predictors of DF cases. Model validation was assessed using Theil's coefficient of inequality, resulting in a value of 0.22. These studies validate the ability of meteorological factors to be practical predictors of DF in both statistical and machine learning approaches.

Deep leaning models have also been applied to DF forecasting, while incorporating climate factors. Node2Vec graph embedding has been incorporated into Support Vector Machine (SVM), Least Absolute Shrinkage and Selection Operator (LASSO), and Artificial Neural Network (ANN) dengue forecast models in China by Liu et al. (2020). This was used to add population flow between dengue outbreak regions as interaction features, in addition to rainfall, lagged temperature, and lagged case count data from the previous four weeks. However, significant improvements in prediction were not observed from the addition of population flow when assessed by a hit rate metric for the prediction of high-risk areas. Moreover, Long Short-Term Memory (LSTM) recurrent neural networks have recently been applied to the prediction of dengue fever cases from climate factors. Xu et al. (2020) investigated the prediction of DF cases in 20 Chinese cities using monthly case data and 9 climate variables (a selection of air pressure, water pressure, temperature, and rainfall measurements), reduced from 15 after filtering for low variance and high correlation to avoid overfitting. LSTM using transfer learning (LSTM-TL) was compared with LSTM, back-propagation neural network (BPNN), Generalised Additive Models, Support Vector Regression (SVR), and Gradient Boosting Machines. The LSTM and LSTM-TL models displayed lower RMSE values for most cities indicating improved predictive accuracy over the other models, with transfer learning rescuing poor predictions in low-incidence areas. Another study used LSTM to predict dengue cases using data from 2002 to 2012 in Kuala Lumpur (Pham et al., 2018). Their LSTM model utilised

a genetic algorithm technique for optimisation of node count, learning rate, regularisation, and batch size parameters. This model was compared with linear regression and decision tree models to assess DF prediction using daily rainfall, temperature, humidity, and windspeed, along with enhanced vegetation index values. LSTM had the highest performance in terms of mean absolute error and RMSE. While the models discussed here show reasonable performance on a case-by-case basis, the lack of a relative error metric such as MAPE means different studies cannot be directly compared.

1.3 Diarrhoea

1.3.1 Aetiology & Burden

Diarrhoeal disease represents a significant burden of childhood malnutrition and mortality globally and in Vietnam. It is the fifth most common cause of death in children under 5 years of age, accounting for approximately 450 thousand deaths annually (Troeger et al., 2018). In Vietnam, an average of 724 thousand cases per year were reported between 1997 and 2016. Clinical diarrhoea is marked by three or more daily loose bowel movements, and can present as acute watery, acute bloody (dysentery), or persistent diarrhoea (≥14 days of illness). Diarrhoeal disease primarily originates from contaminated food and water, where human and animal faeces are a common source (World Health Organisation, 2017). In Vietnam, paediatric diarrhoea is associated with families with low-income, an absence of piped water and toilets, household crowding, low maternal age, and poor education on sanitation (Anders et al., 2015; Nguyen et al., 2006). These factors are not equally weighted throughout the country—36% of households in the northern mountainous regions do not have access to clean water sources.

Here, ethnic minority children under five are 3.5 times more likely to die of preventable causes (UNICEF, 2018). Malnutrition is also a major risk-factor for diarrhoeal disease in children under five years old. Moreover, diarrhoea worsens this malnutrition which can lead to a positive feedback loop causing further bouts of diarrhoea in vulnerable children. Oral rehydration solution is a cheap, effective treatment, and interventions promoting its use have had a significant impact on decreasing global mortality from diarrhoea (Troeger et al., 2018). Prevention measures include exclusive breastfeeding up to six months old, improved access to clean drinking water, improved sanitation education, and rotavirus vaccination (World Health Organisation, 2017). To address limited rotavirus uptake due to high cost and lack of availability, the Vietnamese state-owned company POLYVAC has developed affordable rotavirus vaccines for more equitable access (PATH, 2019).

Diarrhoea due to intestinal infection is caused by a wide range of pathogens, which vary in prevalence by location and season. In the Global Enteric Multicenter Study, which investigated childhood diarrhoea in sub-Saharan Africa and South Asia, rotavirus, *Cryptosporidium*, enterogenic *Escherichia coli* producing heat-stable toxin, and *Shigella* were identified as the main pathogenic causes. Other pathogens were location-dependent, such as *Aeromonas* which was only found in Asia (Kotloff et al., 2013). In Vietnam, rotavirus has similarly been shown to be the most commonly identified pathogen in studies in the North (Isenbarger et al., 2001; Nguyen et al., 2004), in the South (Anders et al., 2015; Thompson et al., 2015a), and throughout the country (Huyen et al., 2018). Differences in other detected pathogens are obscured due to different pathogens being tested for in the aforementioned studies. However, norovirus was the second-most prevalent infection in the southern studies, and *Campylobacter, Salmonella, Shigella, Bacteroides fragilis,* and diarrheagenic *E. coli* were also prevalent in some of the listed studies. The seasonality of diarrhoeal disease in Vietnam varies by pathogen, with

rotavirus infections peaking in autumn and winter while bacterial infections show higher prevalence during the summer months (Nguyen et al., 2006).

1.3.2 Associations between Climate Factors and Diarrhoea

Various studies have investigated the relationships between climate factors and diarrhoea in Vietnam and other countries, though they are less prevalent than those examining DF. Increases in weekly river level have been shown to correlate with diarrhoea rates 0–1 week(s) ahead (Phung et al., 2015a; Thompson et al., 2015b). Previous works have found negative associations between average rainfall and diarrhoeal disease in the same month (Thompson et al., 2015b) or positive associations 0–2 months in advance (Wangdi and Clements, 2017). Total monthly rainfall has been found to correlate with diarrhoea negatively in the same month (Phung et al., 2018) and positively at a 1-month lag (Phung et al., 2015c). Additionally, periods of heavy rainfall, specifically, have been found to precede higher diarrhoea rates by 4–6 days (Phung et al., 2017). Multiple studies have found maximum temperature (Wangdi and Clements, 2017) or average temperature (Phung et al., 2018, 2015c) to be positively correlated with diarrhoea 0–2 months in advance. D'souza et al. (2008) found an overall negative correlation between average temperature and diarrhoea at a 1-month lag in three Australian cities, however this relationship differed across seasons in Brisbane. One study by Onozuka and Hashizume (2011) showed a non-linear relationship, with a 23.2% increase in paediatric infectious diarrhoea cases for every 1°C increase in temperature up to 13°C and an 11.8% decrease for every 1°C increase after that threshold. Studies on relative humidity have been similarly dichotomous, showing negative correlations at 0–2-week lags (D'souza et al., 2008; Phung et al., 2015a; Thompson et al., 2015b) or positive correlations at 0–1 month lags (Phung

et al., 2018, 2015c). The relationships described here highlight climate factors as potential diarrhoea predictors. However, the differing findings within and outside of Vietnam suggest it may be important to select prediction factors and design models at the provincial level to account for these variations.

1.3.3 Diarrhoea Prediction Models

Previous ARIMA-based works have attempted to predict diarrhoeal disease rates from climate factors; a selection of these models is discussed below as well as their results where relative error scores were provided. Kam et al. (2010) examined SARIMA models, comparing univariate models with multivariate SARIMA with exogenous regressors (SARIMAX) models in the prediction of acute diarrhoea patient count in Seoul, South Korea. Exogenous variables were selected by significance during each step of walk-forward validation from patient data, holiday data, and a weather database of average temperature, minimum temperature, maximum temperature, temperature difference, precipitation, wind speed, humidity, and sunshine hours. Climate factors were not significant, however their inclusion reduced average MAPE from 11% to 10%. A multivariate SARIMAX model was implemented by Ali et al. (2013) to forecast cholera-specific diarrhoea cases in Matlab, Bangladesh. Cholera incidence was significantly influenced by minimum temperature and surface sea temperature in the same month, and by surface sea temperature at a 2–month lag. Overall, given the MAPE of 10% obtained in Seoul, ARIMA-based models appear to show reasonably strong predictive accuracies in certain geographic regions. Additionally, climate factors are able to increase model performance.

A small number of other machine learning models have also been applied to the problem of diarrhoea prediction. Fang et al. (2020) compared a random forest model with ARIMA and ARIMAX models in forecasting diarrhoea incidence in Jiangsu Province in China. The random forest model used air pressure, rainfall, and relative humidity as climate factors, while the multivariate ARIMAX model used precipitation. The random forest model had the lowest MAPE of 21%, while the ARIMAX and ARIMA models had values of 28% and 29%, respectively. Sahai et al. (2020) developed a Self-Organising Map, an unsupervised neural network model, to predict diarrhoeal disease two to three weeks in advance in two cities in India, Pune and Nagpur. Minimum temperature, maximum temperature, and rainfall were used as climate parameters, resulting in correlation coefficients of 0.72 in both locations. Lastly, Abdullahi and Nitschke (2021) compared Convolutional Neural Network (CNN), LSTM, and SVM diarrhoea prediction models in nine South African provinces, using eight measures of temperature, precipitation, humidity, air pressure, evaporation, and windspeed as climate predictors. When using real-world data, CNN had the lowest overall RMSE scores, followed by LSTM then SVM. These experiments propose deep learning models as viable diarrhoea forecasting techniques. Moreover, they indicate a potential higher forecasting performance from deep learning models over traditional methods such as ARIMAX.

1.4 Study Aims, Overview, and Contributions

This project aimed to develop accurate prediction models for short-term (one month ahead) and long-term (three months ahead) DF and diarrhoea forecasting based on a diverse range of climate factors in Vietnam. In the national health sector's climate change action plan, the

development of early warning systems for DF and diarrhoea was identified as a priority adaptation measure (Tuyet-Hanh et al., 2018a). The current lack of such systems represents a public health risk and an opportunity for impactful reduction in infectious disease morbidity and mortality in Vietnam. Therefore, a selection of machine learning models for infectious disease forecasting were developed using a dataset of 12 meteorological variables from 1997–2016 in multiple provinces across Vietnam. In the preliminary stages of the project, five traditional statistical and machine learning models (Poisson regression, extreme gradient boosting (XGBoost), SVR, linear kernel SVR (SVR-L), and SARIMA) were developed and compared with four deep learning models (CNN, LSTM, attention mechanism-enhanced LSTM (LSTM-ATT), and Transformer) in 20 provinces across Vietnam. For the second stage of the project, which focused on diarrhoea prediction models, automated hyperparameter optimisation was introduced. This allowed a more systematic approach, and a revisiting of traditional machine learning model following their poor relative performance in DF forecasting. As such, SARIMA and SARIMAX were included to examine their competitiveness with the highest performing models from section one—CNN, LSTM, and LSTM-ATT. Performance was evaluated on the five provinces with the highest median monthly diarrhoea rates, as a proxy for poor clean water and water infrastructure access.

To the best of our knowledge, there have been no deep learning models published for the prediction of DF incidence based on climate factors. Similarly, we have not come across the use of climate-based LSTM-ATT models for DF or diarrhoea forecasting. This study appears to be the first in Vietnam to forecast long-term diarrhoea incidence, and the second to forecast long-term DF incidence (Colón-González et al., 2021). While we focused on diarrhoea prediction in provinces with limited access to clean water, we evaluated our DF models on a

much larger scale than most previous works in Vietnam—20 provinces throughout the northern, central, and southern regions of the country.

2. Materials and Methods

2.1 Data

Disease case data were provided by the Vietnamese National Institute of Hygiene and Epidemiology (NIHE), which included monthly case and death numbers for DF, diarrhoea, and influenza at the provincial level. Disease incidence was provided rather than prevalence, meaning monthly disease case numbers refer only to new cases that were registered that month. Monthly climate data from 1997–2016 consisting of 12 measurements were provided by the Vietnam Institute of Meteorology, Hydrology and Climate Change (IMHEN): total rainfall (mm), highest daily rainfall (mm), number of rainy days, average temperature (°C), absolute minimum temperature (°C), absolute maximum temperature (°C), minimum average temperature (°C), maximum average temperature (°C), average and minimum humidity (%), total evaporation (mm), and total sunshine hours. Additionally, annual population data by province was sourced from the General Statistics Office of Vietnam (2021a) to calculate disease rates from case numbers.

A small selection of statistical tests was applied to adequately describe the data and differences between regions of Vietnam. Normality was tested for each variable using the Shapiro-Wilk test implemented as *scipy.stats.shapiro* in SciPy (version 1.7.1) (Virtanen et al., 2020). Non-

parametric tests were used to investigate differences between regions. The Kruskal-Wallis H-test was used to test for differences in the population medians of disease and climate variables between north, south, and central Vietnam. This was achieved through the *scipy.stats.kruskal* function in SciPy (version 1.7.1) (Virtanen et al., 2020). Following this, post-hoc comparisons were conducted using Dunn's test for pairwise comparisons of mean rank sums, as implemented in the scikit-posthocs (version 0.6.7) function, *scikit_posthocs.posthoc_dunn* (Terpilowski, 2019). Bonferroni correction was applied to significance thresholds to account for multiple hypothesis testing. Significance thresholds were divided by the total number of comparisons between regions, which was 54.

2.2 Data Pre-processing

The climate and disease datasets required pre-processing in the form of imputation, normalisation, and rates calculation. Firstly, both datasets contained some missing values (0–1.36% per variable) in the form of NAs. The datasets also contained 0s, however these were assumed to be true values as there is no way to check, for example, if a given day had 0 cases of dengue fever or if the 0 represents missing data. Missing data was imputed as the minimum value for the same month from the past two years where possible. Missing values in the first 12 months were imputed as 0, missing values in months 13–24 were imputed as the value for the same month from the previous year, and missing values in all other months were imputed as the minimum value of the same month from the previous two years. Secondly, the multivariate models (i.e., Poisson regression, SVR, SVR-L, XGBoost, SARIMAX, CNN, LSTM, LSTM-ATT, Transformer) required normalisation of the data to ensure all prediction

factors were equally weighted. To achieve this, data was scaled on the training set in the range of 0–1 using MinMaxScaler from Scikit-learn (version 0.24.2) (Pedregosa et al., 2011). Finally, disease rates per 100,000 population were calculated from the case data using annual population data to use for analysis and forecasting, in order to account for population size.

2.3 Visualisation

Data were visualised through geospatial mapping as well as standard plots. Geospatial mapping of median disease incidence was achieved through GeoPandas (version 0.9.0) (Jordahl et al., 2020). After amending differences in province names, the project data was merged with a JSON file of Vietnam containing the coordinates and multipolygon data for each province. Next, the GeoSeries Coordinate Reference System was converted to Web Mercator to allow the addition of CARTO Positron basemaps (CARTO, 2019). The map plots were then generated by Matplotlib (version 3.4.3) (Hunter, 2007). Boxplots, bar plots, and time series plots were generated using Matplotlib (version 3.4.3) (Hunter, 2007) and Seaborn (version 0.11.2) (Waskom, 2021).

2.4 Prediction Models

2.4.1 Poisson Regression

Poisson regression models are a type of generalised linear model (GLM). GLMs are a class of statistical models which allow the response variable to follow a non-normal distribution—a Poisson distribution is assumed in the case of Poisson regression. Poisson regression models also assume the log of the response variable mean can be modelled by linear independent variables, and is equal to its variance. In this study, the primary model parameter (λ) is the average monthly disease incidence. Modelling the log of λ avoids violating the equal variance assumption and allows values from negative to positive infinity. The Poisson regression can be represented mathematically as shown below, where x_t^i terms are covariates such as previous disease rates or climate factors and β_i terms are coefficients (Roback and Legler, 2021, sec. 4.2).

$$\log \lambda_t = \sum_{i=1}^{n} \beta_i x_t^i + \beta_0$$

2.4.2 Extreme Gradient Boosting (XGBoost)

XGBoost is an implementation of gradient boosting machines, models based on decision tree ensembles. These ensembles are built upon collections of classification and regression trees (CART). A single CART splits datapoints into leaves to determine, in the context of this paper, disease incidence based on previous rates or climate factors. However, one CART is

insufficient to accurately predict disease incidence, so many trees are combined to form an ensemble where prediction weights of the different tree leaves are summed. XGBoost models are trained to minimise an objective function consisting of a loss function (mean squared error) and a regularisation function, which limits overfitting. The models are trained in an additive manner, where one level of the tree is optimised at a time. In practice, this means one leaf is split into two further leaves to see whether the objective function is improved, and many such splits are considered to determine the best possible split (Chen and Guestrin, 2016).

2.4.3 Support Vector Regression (SVR)

SVR is a regression-generalised implementation of SVMs. SVMs aim to categorise datapoints (vectors) by separating them with a hyperplane that maximises the distance (ε) between the resulting groups and the hyperplane. In many cases, however, a linear separator cannot accurately split up the datapoints into the respective groups, so inputs are mapped into higher-dimensional space with a kernel function. This allows separation of the datapoints with hyperplanes in higher dimensions. SVR introduces an ε-tube around the regression line. Support vectors in SVR are the training datapoints outside of this tube. The optimisation function aims to find a regression function that keeps datapoints within the ε-threshold while keeping model complexity low. Importantly, datapoints are not penalised for errors less than ε (i.e., datapoints within the tube) (Awad and Khanna, 2015). SVR models can be implemented with several different kernel options. For this project, both radial basis and linear kernels were implemented in SVR and SVR-L models, respectively.

2.4.4 Seasonal Autoregressive Integrated Moving Average (SARIMA) Models

SARIMA and SARIMAX are developments on the ARIMA time series model, which fits data based on an autoregressive (AR) term, a differencing term, and a moving average (MA) term. As the names suggest, SARIMA(X) models add seasonal terms. SARIMA parameters take the form [(p, d, q) (P, D, Q, s) (trend)], where (p, d, q) are the AR, differencing, and MA terms; (P, D, Q, s) are the seasonal AR, differencing, MA, and interval terms; and trend refers to an optional parameter to account for linear and/or continuous trends. The univariate SARIMA equation is below, where a time series y_t has an error term u_t and a η_t term, in the case of measurement error or pure regression models when p and q are set to 0. In the second line of the equation, $\phi_p(L)$ and $\tilde{\theta}_P(L^s)$ refer to the non-seasonal and seasonal autoregressive lag polynomials, respectively. Similarly, Δ^d and $\Delta_s^D u_t$ refer to the non-seasonal and seasonal differencing. On the right-hand side, $A(t)$ is the trend polynomial with intercept, $\theta_q(L)$ is the non-seasonal moving average lag polynomial and $\tilde{\theta}_Q(L^s)$ is the seasonal equivalent, and ζ_t is the noise component (Perktold et al., 2021a, 2021b).

$$y_t = u_t + \eta_t$$

$$\phi_p(L)\tilde{\theta}_P(L^s)\Delta^d\Delta_s^D u_t = A(t) + \theta_q(L)\tilde{\theta}_Q(L^s)\zeta_t$$

SARIMAX models are multivariate due to the addition of exogenous (X) regressors, and take the form of regression models with SARIMA errors (Perktold et al., 2021a). This allows standard regression interpretation of covariate coefficients, which is not possible for the alternative model formation where covariates are simply added to the right-hand side of the SARIMA regression equation (Hyndman, 2010). The multivariate SARIMAX equation is

below, where there is an addition of n exogenous regressor terms x_t^i with coefficients β_i (Perktold et al., 2021a, 2021b). Exogenous variables at different time lags are considered as separate terms.

$$y_t = \sum_{i=1}^{n} \beta_i x_t^i + u_t$$

$$\phi_p(L)\tilde{\theta}_P(L^s)\Delta^d\Delta_s^D u_t = A(t) + \theta_q(L)\tilde{\theta}_Q(L^s)\zeta_t$$

As the variables were scaled between 0 and 1 individually, the β_i coefficient represents the mean unit change in y (i.e., scaled disease incidence) for a 1-unit change in x^i (i.e., scaled exogenous factor). To find the unscaled change, the values were inverse transformed using the *inverse_transform* function from Scikit-learn (version 0.24.2) MinMaxScaler (Pedregosa et al., 2011). Finally, Bonferroni corrections were applied to adjust for multiple hypothesis testing. Significance thresholds were divided by the number of exogenous factors per province (6). Below, $\widehat{\beta_i}$ is the mean unit change in disease incidence for a 1-unit change in an exogenous factor $\widehat{x_i}$; $sc_y.inverse\_transform$ and $sc_{x^i}.inverse\_transform$ refer to the inverse scaling transformations for disease incidence and exogenous factor, respectively.

$$\widehat{\beta_i} = \frac{sc_y.inverse\_transform\ (1)}{sc_{x^i}.inverse\_transform(\beta_i)}$$

2.4.5 Convolutional Neural Network (CNN)

CNNs are a form of ANNs, a class of machine learning models based on many interconnected processors activated by the environment or other neurons which ultimately accumulate in a desired outcome such as time series prediction. ANNs consist of an input layer of nodes (neurons), layers of hidden nodes, and an output layer of nodes. As a model learns, the weighted connections between nodes are updated in a linear or non-linear fashion to tune the activation of the network (Schmidhuber, 2015). CNNs reduce the parameters needed in ANNs, implementing a convolution operation to extract key features from the data before pooling to reduce complexity. This can be repeated with further convolutional and pooling layers before feeding the output into a fully connected layer, comparable to a regular ANN (Albawi et al., 2017). The CNN model used in this study was made up of a 1-dimensional (1D) convolution layer, a 1D max pooling layer, and one fully connected layer which outputs predictions for up to three months ahead (Figure 1).

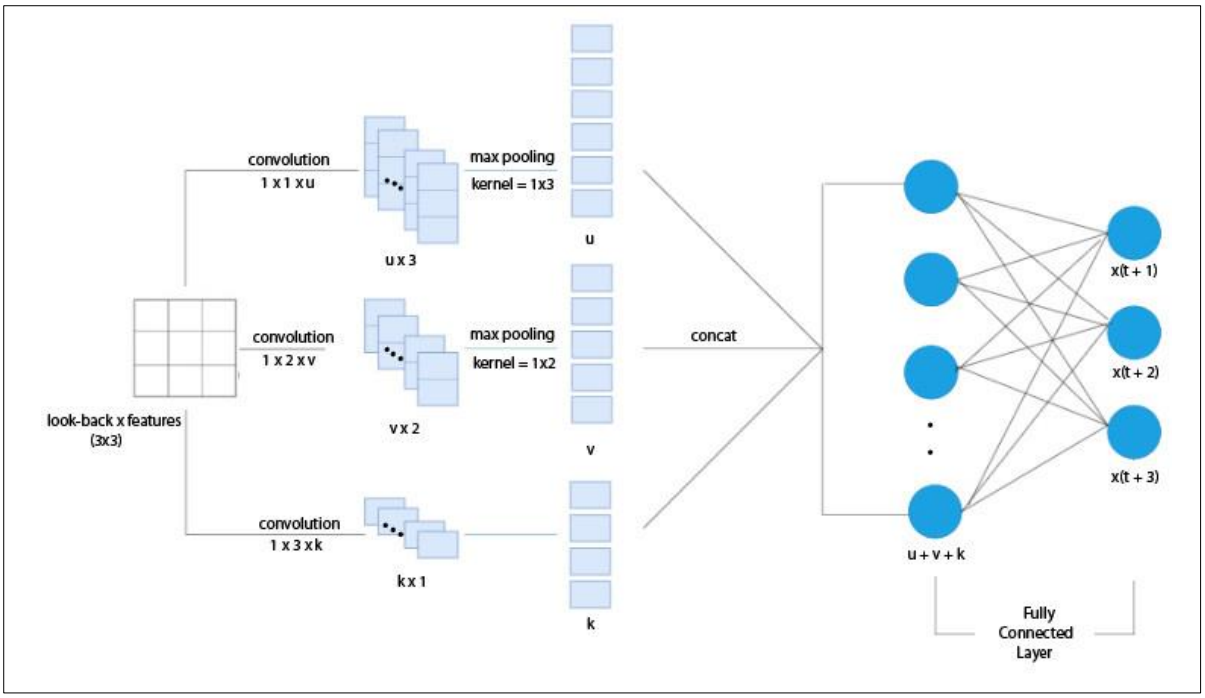

Figure 1: Convolutional Neural Network architecture. The model consists of a 1-dimensional (1D) convolution layer, a 1D, max pooling layer, and a final fully connected layer which results in predictions for disease incidence one, two, and three months ahead.

2.4.6 Long Short-Term Memory (LSTM)

LSTM models are a variant of recurrent neural networks (RNNs). In contrast to feedforward neural networks which can only pass on information in one direction, RNNs can also recurrently feed outputs back into the model as inputs. RNNs, however, struggle to learn longer sequences (Graves and Schmidhuber, 2005). The LSTM architecture overcomes this, as well as the common deep neural network problem of vanishing or exploding gradients caused by cumulative backpropagation error signals (Schmidhuber, 2015). LSTMs implement memory blocks in place of regular RNN hidden layers. These memory blocks have a cell state in addition to an input gate, a forget gate, and an output gate. The specific LSTM model used in this research consists of one memory block for each time point in the lookback window (Figure

2). In most cases, this is the previous three months. The results from the final hidden state are fed into a fully connected layer, which outputs predictions for the next three months.

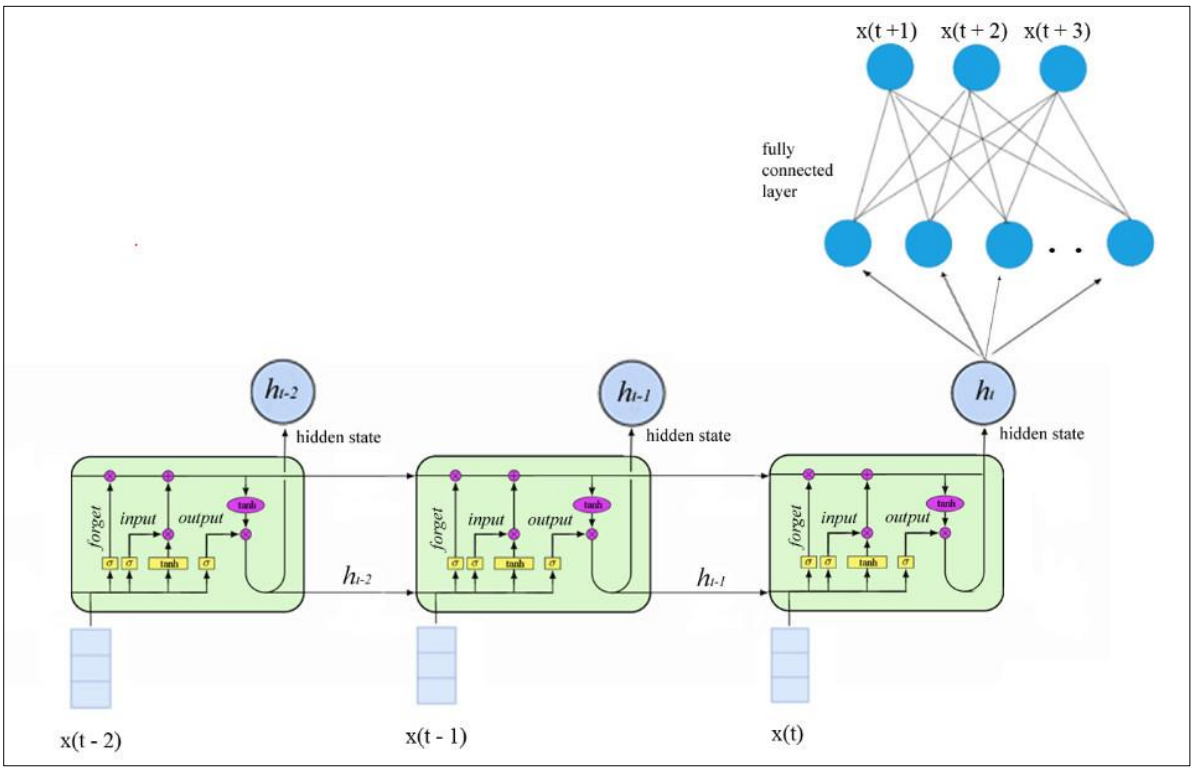

Figure 2: Long Short-Term Memory model architecture. For a lookback window of three months, the model is made up of three memory cells which ultimately feed through a fully connected layer to output predictions for the next three months.

2.4.7 Attention Mechanism-enhanced Long Short-Term Memory (LSTM-ATT)

Attention mechanisms in machine translation were initially developed to reduce the loss of information between sequence steps, by allowing models to focus on the most important parts of input data (Bahdanau et al., 2016). In LSTMs, this is achieved by generating outputs from each memory cell hidden state, and has been shown to improve model performance when

dealing with long sequences (Luong et al., 2015). An attention layer was added after the LSTM network, based on the global attention mechanism introduced by Luong et al. (Luong et al., 2015) (Figure 3). In brief, the LSTM-ATT model takes a target hidden state from the top layer of the LSTM (i.e., the most recent month) and combines it with a context vector developed from the hidden states from all lookback window months. This produces an attentional hidden state, which is then fed into a fully connected layer to output disease incidence predictions for the following three months.

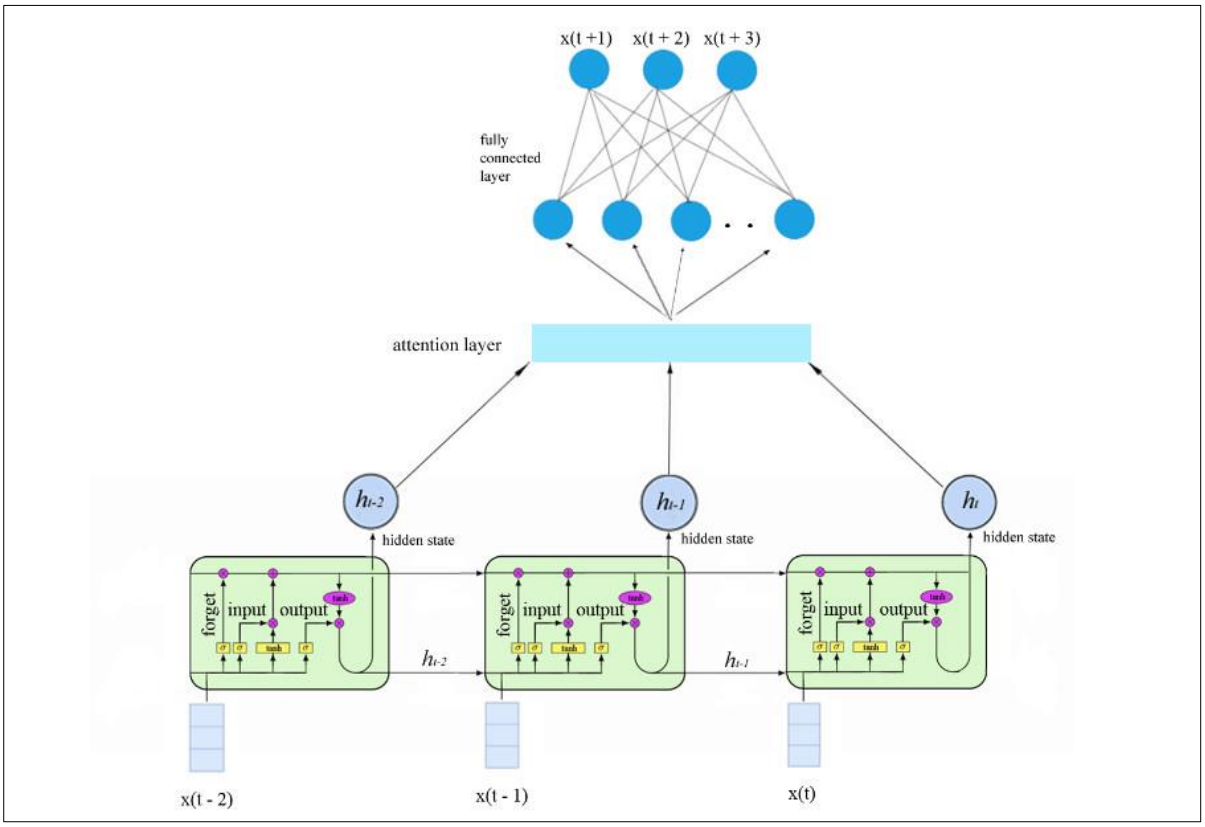

Figure 3: Long Short-Term Memory with Attention Mechanism model architecture. For a lookback window of three months, the model is made up of three memory cells. The hidden states of these cells are fed into an attention layer before the fully connected layer, which outputs predictions for the next three months.

2.4.8 Transformer

The transformer is another deep learning model that exploits attention mechanisms, which was recently developed and shown to have strong performance in natural language processing (Vaswani et al., 2017). However, in contrast to the previously described models, it does not use recurrence or convolutions. The transformer model relies solely on attention and is able to process input data not in order, which allows improved parallelisation and reduced training times. Its model architecture includes encoder and decoder layers which contain multi-head self-attention mechanisms which flow into feed-forward networks (Figure 4) (Vaswani et al., 2017).

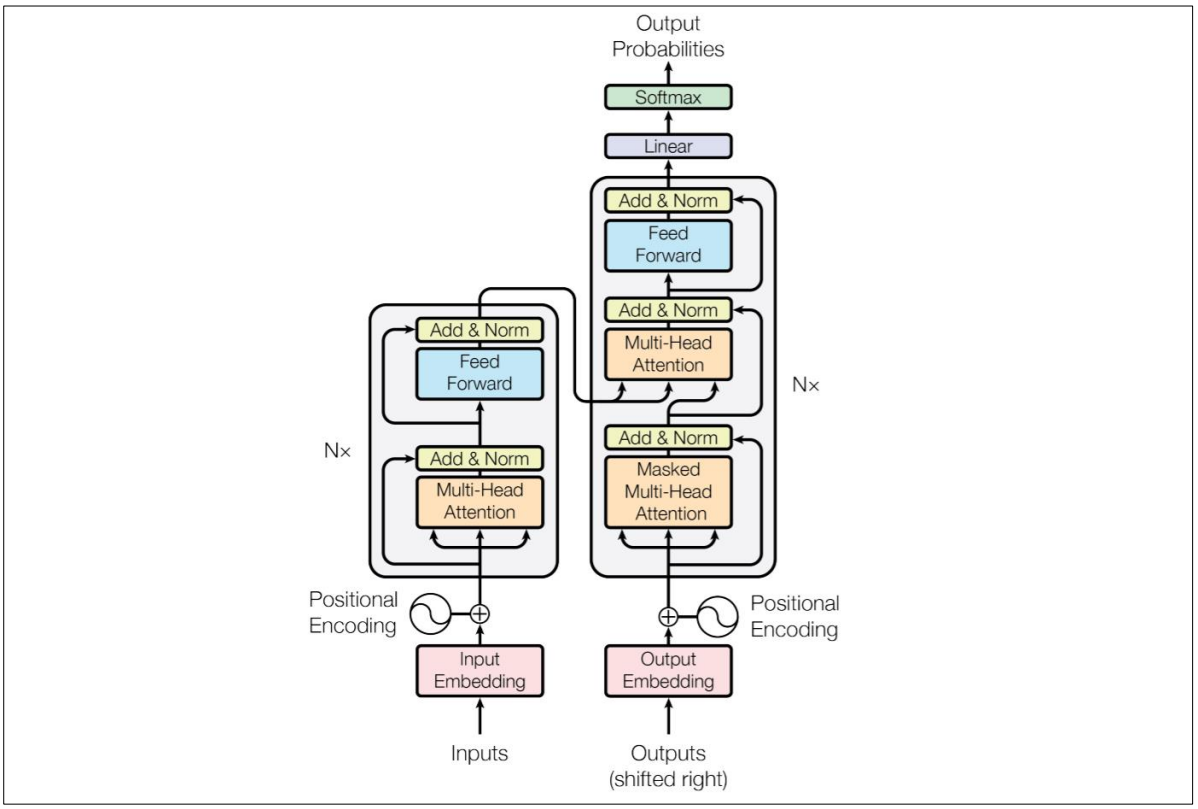

Figure 4: Transformer model architecture. Disease incidence and climate data is fed into the model as inputs, and disease incidence predictions for the following three months the final outputs Figure from Vaswani et al. (2017).

2.5 Hyperparameter Optimisation & Model Implementation

2.5.1 Traditional Models

Poisson regression, SVR, and SVR-L models were implemented in Scikit-learn (version 0.24.2) (Pedregosa et al., 2011), while XGBoost models were implemented in the XGBoost Python package (version 1.5.0) (Chen and Guestrin, 2016). For the Poisson regression models, *alpha* was set to *1e-15*, and *max_iter* to *1e6*. For XGBoost models, default parameters were used. For SVR, the following parameters were used: *kernel = "rbf", C* = 100, *gamm*a = *"auto",* epsilon = 0.1. For SVR-L, the following parameters were used: *kernel = "linear", C = 100, gamma = "auto".*

Univariate SARIMA and multivariate SARIMAX models were implemented using the SARIMAX model from the statsmodels (v0.12.2) Python library (Seabold and Perktold, 2010). Initially in the project, trial and error was used for hyperparameter selection for optimisation of the SARIMA DF models. Default parameters were used with the exception of *enforce_stationarity* and *enforce_invertibility*, which were set to false. Firstly, for each model, the time series was decomposed into seasonal, trend, and residual components which revealed the seasonality (m) and trend parameters. Secondly, the augmented Dickey-Fuller test was used to test if the data was stationary or required differencing (d/D). Thirdly, autocorrelation function (ACF) and partial ACF (PACF) plots were examined to help determine moving average (q/Q) and autoregressive (p/P) terms before and after differencing.

For the second section of the project which focused on diarrhoeal disease forecasting, more systematic methods were used for hyperparameter selection where practical. In cases where it

was impractical, trial and error were used. Recursive feature selection was performed with a random forest regressor from Scikit-learn (version 0.24.2) (Pedregosa et al., 2011) to select two exogenous predictors. Exploratory analysis on the training data revealed a potential correlation between diarrhoea and influenza rates. To accommodate this, influenza rates were added to the climate dataset for recursive feature selection. The seasonal parameter (m) was set to 12 as the data showed clear annual seasonality. All other hyperparameters were chosen using Bayesian model-based optimisation. This was implemented with a Tree-structured Parzen Estimator (TPE) in Optuna (version 2.8.0) (Akiba et al., 2019) which aimed to minimise RMSE. Initially, exhaustive grid-search was attempted but was too computationally intensive to run in a practical amount of time, even in parallel on a high-performance computing cluster. ACF, PACF, and augmented Dickey-Fuller functions were still examined before and after differencing to manually check the fit of the parameters to the data. Similarly, the decomposed time series was examined to check seasonality and trend before and after differencing.

2.5.2 Deep Learning Models

Pytorch (version 1.8.1) (Paszke et al., 2019) and Scikit-learn (version 0.24.2) (Pedregosa et al., 2011) were used to create and implement the deep learning models (CNN, LSTM, LSTM-ATT, Transformer). As for the SARIMA models, trial and error methods were used initially in the project for deep learning model DF predictions. This included experimenting with lookback windows of 1–18 months, resulting in an optimum window length of 3 months. Across all models, common hyperparameters were set as follows: *batch size* = 16, *learning rate* = $1e^{-3}$, *dropout* = 0.1, and *epochs* = 300. The CNN model had the following hyperparameters: *number of layers* = 1, *number of filters* = 100, and *kernel sizes* = (1, 3), (2, 3), and (3, 3). For the LSTM,

LSTM-ATT, and Transformer models, the numbers of layers and hidden sizes were tuned specifically for each province (Table S1). For both DF and diarrhoea predictions, the numbers of features used in the models were chosen using trial and error, and Adaptive Movement Estimation (Adam) was used as a gradient descent optimisation algorithm due to consistent performance and popularity described in the literature (Okewu et al., 2019). Likewise, two exogenous predictors and *batch size = 16* were used for both disease predictions. For the diarrhoea forecasting models, the TPE in Optuna (version 2.8.0) (Akiba et al., 2019) selected province-specific values for *lookback window, epochs, learning rates, hidden sizes,* and *numbers of layers* for LSTM and LSTM-ATT models. *Epochs, learning rates, number of filters,* and *dropout rates* were similarly optimised for CNN models. The specific values for each province are available in the supplementary materials (Table S2).

2.6 Performance Evaluation

2.6.1 Forecasting Evaluation

The 20 years of data was split into a 14 year training set (1997–2010), a three year validation set (2011–2013), and a three year testing set (2014-2016). The models were trained on the training data and evaluated on the validation set to optimise hyperparameters. Then, the tuned models were evaluated on the testing set to compare model performance using RMSE, while MAE and MAPE were also provided for further evaluation of the models. RMSE calculates the square root of the mean of the squared prediction errors between predicted and actual values (i.e., the standard deviation of the prediction errors). RMSE penalises larger errors greater than

smaller errors. Its formula is shown below, where y_i is an actual value and $\hat{y}_i$ is the corresponding predicted value:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

MAE, in contrast, weights errors equally by calculating the mean of the absolute differences between forecasted and actual values. The MAE formula is provided below, where y_i is an actual value and $\hat{y}_i$ is the corresponding predicted value:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

MAPE is defined as the mean of absolute errors between predicted and corresponding actual values divided by actual values. As MAPE is a relative error, model accuracy for different provinces can be directly compared. Where y_i is an actual value and $\hat{y}_i$ is the corresponding predicted value, MAPE is calculated as follows:

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| * 100\%$$

2.6.2 Outbreak Evaluation

The ability of LSTM-ATT to forecast disease outbreaks was assessed by calculating accuracy, precision, sensitivity, and specificity for each province's test set. Outbreak months were defined as those with case numbers greater than one standard deviation above the average of disease incidence for that month. This method was based on previous works (Brady et al., 2015;

Cheng et al., 2020). Firstly, accuracy represents the proportion of predictions that were correct for a province. Secondly, precision is defined as the number of correctly predicted outbreaks relative to the total number of actual outbreaks. Thirdly, sensitivity represents the proportion of actual outbreaks that were correctly detected. Finally, specificity is defined as the proportion of normal (non-outbreak) months that were correctly detected.

$$Accuracy = \frac{Correct\ Predictions\ (Outbreak\ or\ Normal\ Months)}{Total\ Predictions}$$

$$Precision = \frac{Correct\ Predictions\ (Outbreak\ Months)}{Total\ Predicted\ Outbreak\ Months}$$

$$Sensitivity = \frac{Correct\ Predictions\ (Outbreak\ Months)}{Total\ Outbreak\ Months}$$

$$Specificity = \frac{Correct\ Predictions\ (Normal\ Months)}{Total\ Normal\ Months}$$

3. Results

3.1 Descriptive and Statistical Analyses of Datasets

There were 1,618,767 cases of DF in Vietnam between 1997 and 2016. DF cases were not normally distributed, nor were any of the other variables ($p < 0.001$, Table S3). During this time period, there was a median monthly DF rate of 0.702 per 100,000 population. Incidence and death rates of DF were highest in the months of June to October. Median incidence rates of dengue fever were highest in the Southern provinces ($p < 0.001$). Additionally, median

dengue fever rates were significantly higher in central Vietnam than in northern Vietnam ($p < 0.001$) (Figure 5).

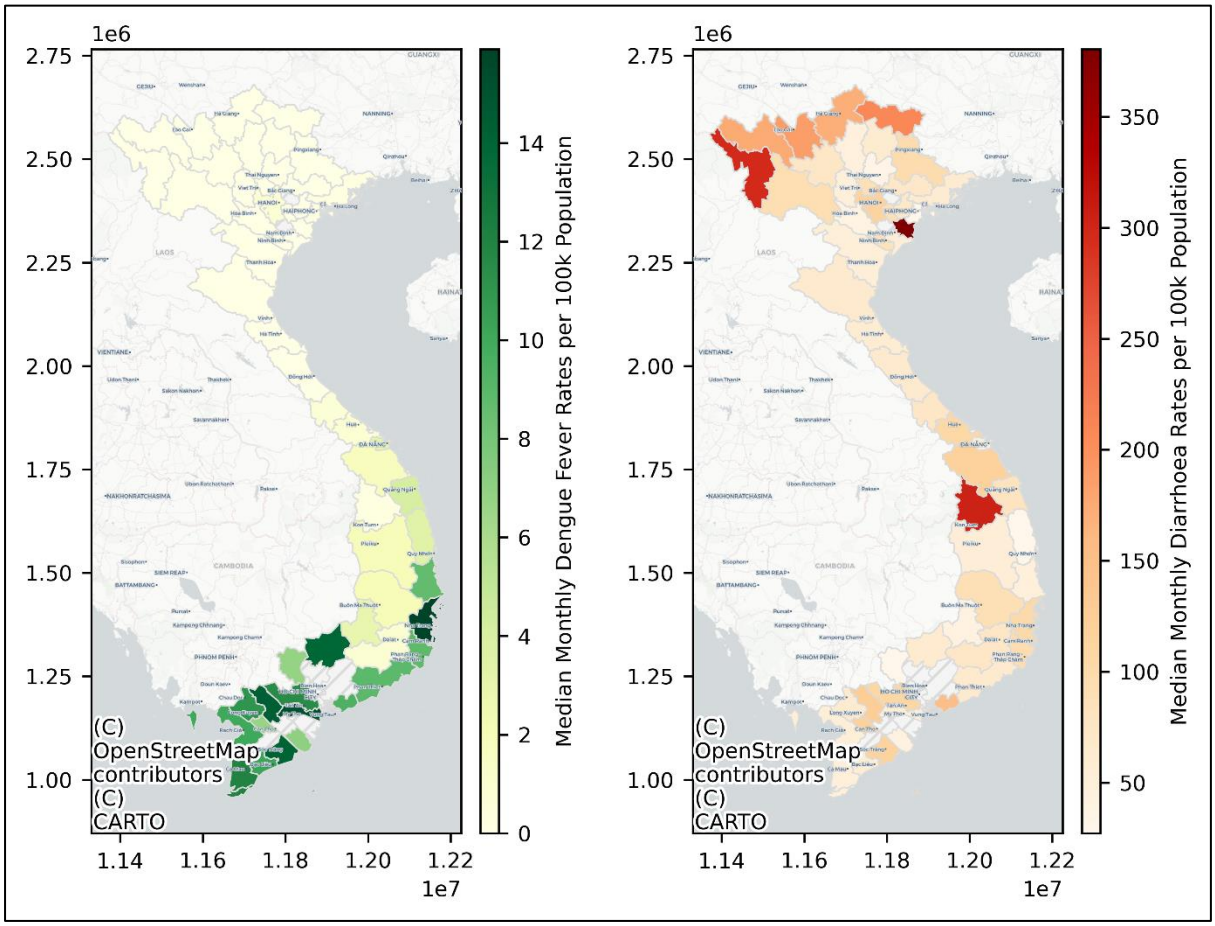

Figure 5: Median monthly incidence rates of dengue fever and diarrhoea in Vietnam. Axes are shown in meters. Basemaps were obtained from CARTO (2019).

The visual spread of DF rates per province suggests large differences between provinces. Some provinces experienced very few infections at all from 1997–2016, as shown by interquartile ranges (IQRs) hovering above zero (e.g., Bắc Cạn, Cao Bằng, Hà Giang). Others had either consistently higher median DF rates, such as in An Giang, Bình Phước, and Phú Yên, or a larger spread of IQR values covering rates from zero to the hundreds, as seen in Kon Tum, Quảng Bình, and Quảng Trị. However, there were consistently many outliers plotted, indicating

data points outside 1.5x the IQR, representing the presence of outbreak-like spikes in cases relative to normal rates (Figure 6).

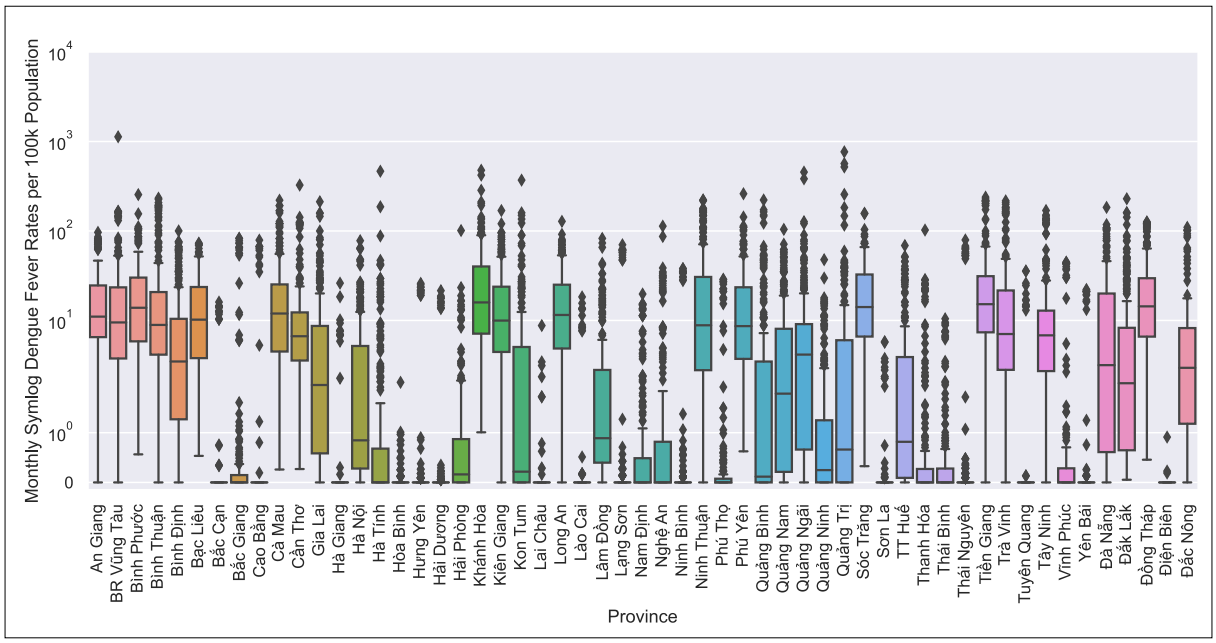

Figure 6: Distribution of dengue fever incidence rates per province in Vietnam. Box and whisker plots are drawn for 55 of Vietnam's 63 provinces, where boxes indicate interquartile ranges and points outside 1.5 times the IQR are marked as outliers.

There were 14,476,100 cases of diarrhoea between 1997 and 2016, with a median monthly diarrhoea rate of 97.8 cases per 100,000 population. The highest median rates of diarrhoea were mostly in the mountainous provinces—the upper northern provinces (e.g., Lào Cai, Điện Biên, and Cao Bằng) and Kon Tum in the Central Highlands—along with Thái Bình in the Red River Delta (Figure 1). Diarrhoea rates were significantly higher in the North than in southern or central Vietnam ($p < 0.001$). There were fewer outliers plotted in comparison to DF, though they are still present for many provinces showing outbreak-like spikes. In contrast to DF, all provinces had IQRs above zero, suggesting a more constant presence of case-registered diarrhoeal disease in Vietnam (Figure 7). This is supported by the data, as there were only 288

months throughout all provinces without diarrhoea cases registered, as opposed to 4277 for DF.

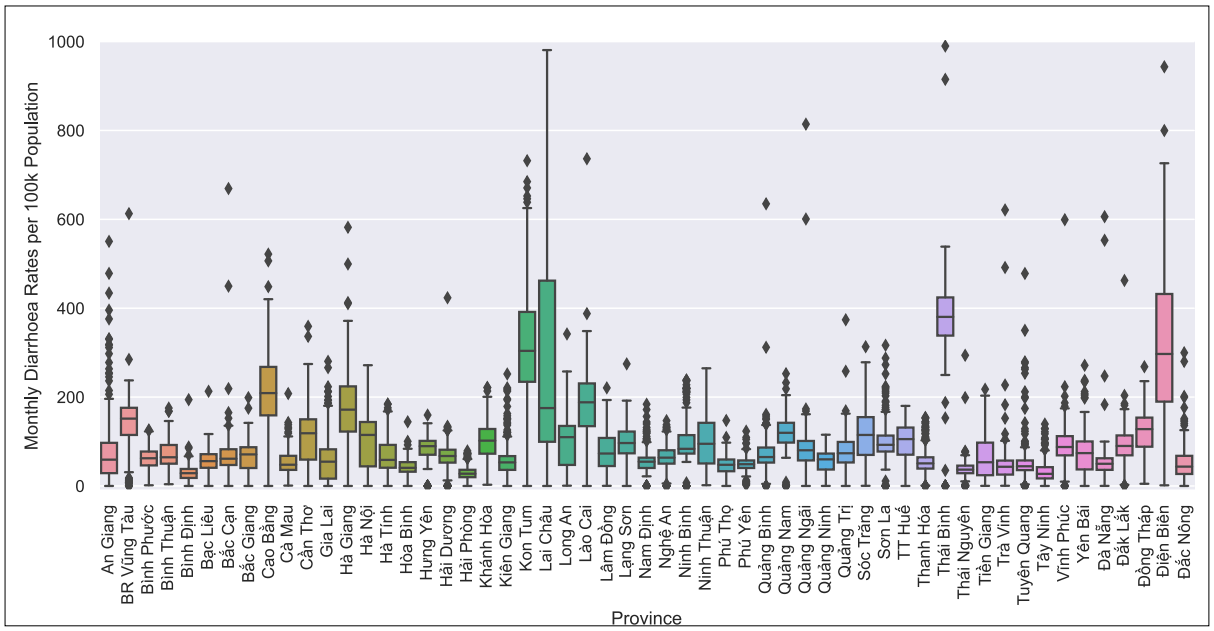

Figure 7: Distribution of diarrhoea incidence rates per province in Vietnam. Box and whisker plots are drawn for 55 of Vietnam's 63 provinces, where boxes indicate interquartile ranges and points outside 1.5 times the IQR are marked as outliers.

Absolute temperatures in Vietnam ranged between -4.2°C to 41.1°C in the period of 1997–2016, with a median average temperature of 26.0°C throughout the country (Table 1). The median average temperature significantly increased while moving down the regions, from 23.7°C in the North, to 26.5°C in central Vietnam, to 27.4°C in the South ($p < 0.001$). The number of rainy days ranged greatly from 0 to 31 days, which was consistent throughout the northern, central, and southern regions. Median values for monthly total rainfall were highest in the South (137mm), followed by the central region (101.0mm), and then the North (98.0mm). However, these differences were not significant ($p > 0.05$). Average humidity ranged from 50.0–99.0%, and median values were marginally higher in the North at 84.0%

compared to 83% in central ($p < 0.05$) and southern Vietnam ($p < 0.001$). In contrast, median total evaporation rose from 67.3mm to 82.0mm to 84.2mm moving from northern to central to southern Vietnam, with significant differences between each region ($p < 0.001$). Similarly, the median number of monthly sunshine hours also increased moving southwards down the regions, from 130.3 to 172.7 to 203 ($p < 0.001$). Throughout the country, the number of sunshine hours varied from 0 to 321.1 per month. Detailed statistical results are available in the supplementary materials for Kruskal-Wallis (Table S4) and Dunn tests (Table S5).

Table 1: Climate data. Provided by the Vietnam Institute of Meteorology, Hydrology and Climate Change (IMHEN). All variables relate to monthly measurements. S.D. = standard deviation, IQR = interquartile range.

Climate Factor	Mean	Median	S.D.	Min	IQR	Max
Average temperature (°C)	24.6	26.0	4.1	3.8	22.1 – 27.7	31.8
Maximum average temperature (°C)	29.1	30.5	4.4	5.7	26.7 – 32.3	38.2
Minimum average temperature (°C)	21.8	23.1	4.1	2.6	19.3 – 24.9	28.5
Absolute maximum temperature (°C)	33.0	33.4	3.5	13.5	31.2 – 35.4	41.1
Absolute minimum temperature (°C)	18.5	20.3	5.4	-4.2	15.0 – 23.0	26.7
Total rainfall (mm)	160.8	111.5	177.3	0.0	30.9 – 238.0	3207.0
Highest daily rainfall (mm)	47.2	37.0	48.9	0.0	14 – 64	993.0
Number of rainy days	10.4	10.0	8.1	0.0	3.0 – 16.8	31.0
Average humidity (%)	82.8	83.3	4.9	50.0	80.0 – 86.2	99.0
Minimum humidity (%)	49.3	50.0	9.6	11.0	43.0 – 56.0	85.0
Total Evaporation (mm)	79.4	73.5	31.7	1.0	58.1 – 95.9	245.7
Total sunshine hours	159.1	160.7	65.5	0.0	115 – 206.0	321.1

3.2 Dengue Fever

3.2.1 Predicting Dengue Fever One Month in Advance

In general, the deep learning models outperformed the traditional models, as observed by a clear shift from green towards red in the colour-coded RMSE results (Table 2). Error scores were colour-coded to show the range of RMSE values for each province rather than colour-coding them across all provinces, because RMSEs as a measure of model performance are only comparable where the observed DF rate is the same. The traditional models performed worse than LSTM-ATT in all provinces, LSTM in all but one province (Quảng Ngãi), CNN in all but three provinces (Hải Phòng, Quảng Ninh, and Nam Định), and the Transformer in all but three provinces (Hải Phòng, Nam Định, and Bình Thuận). MAEs were similarly higher overall in the traditional models compared to the deep learning models (Table S6). In the multivariate models, the climate factors selected by random forest regression were most commonly measures of rainfall and temperature, chosen 16 and 15 times, respectively. There were also 4 uses of sunshine hours, 3 uses of evaporation, and 2 uses of humidity (Table S7).

Table 2: Root mean square errors for all prediction models in 20 Vietnamese provinces. Values are colour-coded for each province separately from the lowest value (darker green) to the median value (yellow) to the highest value (darker red). LSTM = long short-term memory. LSTM-ATT = attention mechanism-enhanced LSTM. CNN = convolutional neural network. Poisson = Poisson regression. XGB = Extreme Gradient Boosting. SVR = Support Vector Regressor with Radial Basis Kernel. SVR-L = Support Vector Regressor with Linear Kernel. SARIMA = Seasonal Autoregressive Integrated Moving Average.

Province	Root Mean Square Error for Each Model								
	LSTM	LSTM-ATT	CNN	TF	Poisson	XGB	SVR	SVR-L	SARIMA
Hà Nội	7.999	6.630	9.180	11.301	17.162	13.382	16.689	16.878	18.144
Hải Phòng	0.464	0.529	0.757	0.748	0.934	0.657	6.073	7.938	2.594
Quảng Ninh	1.010	0.961	1.953	0.930	1.577	1.277	3.384	4.072	1.175
Nam Định	0.783	0.797	0.974	1.008	0.939	1.156	1.454	1.578	0.933
Thái Bình	0.627	0.597	0.598	0.661	0.688	0.738	0.781	0.878	0.676
Quảng Nam	7.382	6.696	6.890	12.678	13.504	11.990	13.969	15.434	16.448
Quảng Ngãi	9.288	8.080	8.874	8.861	11.113	9.096	27.721	37.677	10.181
Phú Yên	9.187	9.544	9.766	12.544	19.278	16.209	19.329	20.562	20.628
Ninh Thuận	5.064	3.959	5.140	8.743	17.260	24.833	20.274	12.441	9.027
Bình Thuận	8.364	8.826	8.259	12.031	12.949	10.302	13.880	14.512	10.120
Tây Ninh	5.123	3.854	6.538	6.500	7.350	9.395	7.213	9.450	6.600
Bình Phước	6.577	7.466	9.063	9.649	14.796	12.574	17.746	17.507	21.731
An Giang	5.699	3.907	3.860	5.461	9.502	8.672	7.777	7.954	10.504
Tiền Giang	4.415	4.098	7.912	5.620	18.336	17.611	14.648	16.247	13.550
Cần Thơ	3.119	2.228	3.997	4.866	8.689	6.595	18.503	27.518	9.349
Trà Vinh	4.462	3.891	4.820	4.482	12.442	13.630	14.752	14.289	10.129
Kiên Giang	2.460	2.976	4.448	3.892	16.070	16.809	16.093	16.455	5.079
Sóc Trăng	6.192	5.887	3.725	4.389	12.671	13.908	12.227	11.946	42.093
Bạc Liêu	3.429	2.652	2.379	2.891	12.324	11.841	10.035	9.584	23.812
Cà Mau	4.490	4.110	5.499	9.043	14.720	20.489	15.279	15.974	17.736

One-step Forecasting

The results generated by the most competitive deep learning models—CNN, LSTM, and LSTM-ATT—were plotted to visualise their respective accuracies. The results from the transformer and traditional models were not graphed due to notably worse performance and to avoid overplotting. Predicted values were graphed against actual values for the test set covering January 2014 to December 2016. Plots are provided for six of the twenty provinces (Figure 8).

In the representative plots, all three models appear to perform well with LSTM-ATT predictions showing slightly stronger adherence to the observed incidence rates. There are some exceptions to this, such as the CNN model's better coverage of the outbreak peaking at month 25 in Quảng Nam.

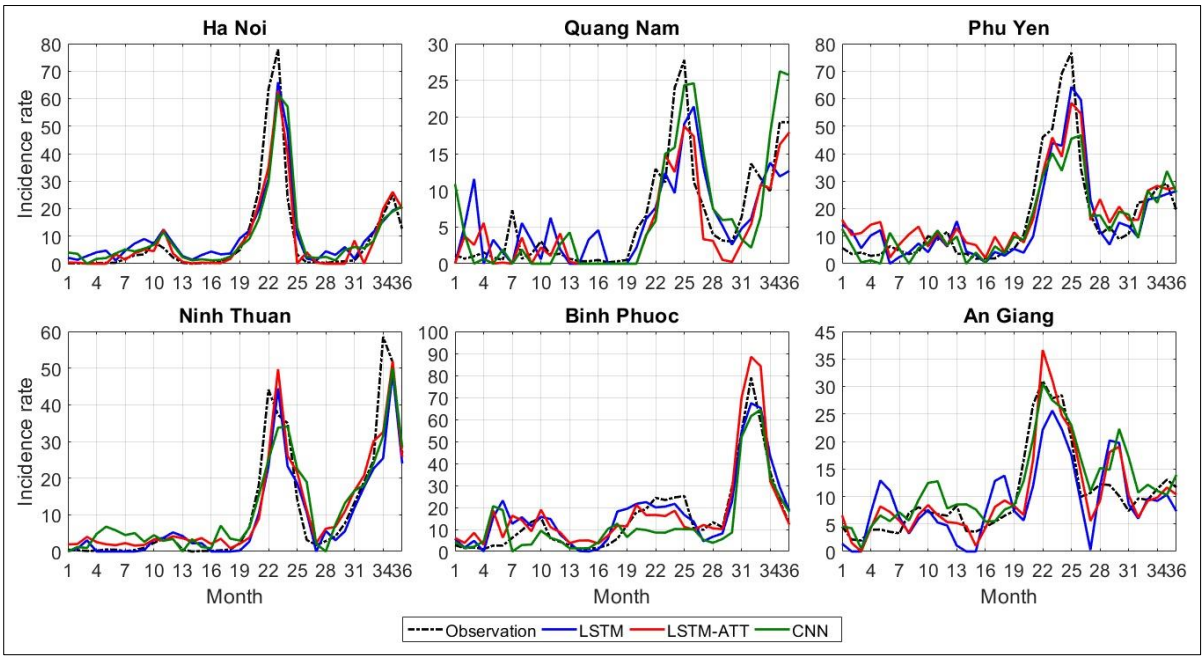

Figure 8: One-month ahead dengue fever predictions for six provinces in Vietnam. Observation refers to the real rates of dengue fever incidence in the test set from 2014–2016. LSTM = Long Short-Term Memory, LSTM-ATT = attention mechanism-enhanced LSTM, CNN = Convolutional Neural Network.

Error metrics were plotted for the deep learning models to provide a clearer picture of respective model performance (Figure 9). In general, the error metrics confirm the differences observed in the plots. LSTM-ATT, LSTM, CNN, and Transformer had the lowest RMSE values for 10, 5, 4, and 1 province(s), respectively. Similarly, LSTM-ATT, LSTM, CNN, and Transformer had the lowest respective MAE values for 8, 5, 5, and 2 provinces. The attention mechanism conferred an overall improvement to the base LSTM model, with decreased

RMSEs in 14/20 provinces and decreased MAEs in 13/20 provinces. MAPEs were inflated by the presence of many low and zero values for observed rates. Even for the highest performing models, MAPE still had some extremely high values ranging from 38.4% in An Giang (LSTM-ATT) to 3.90 x 10^{-16} in Kon Tum (CNN) (Table S8).

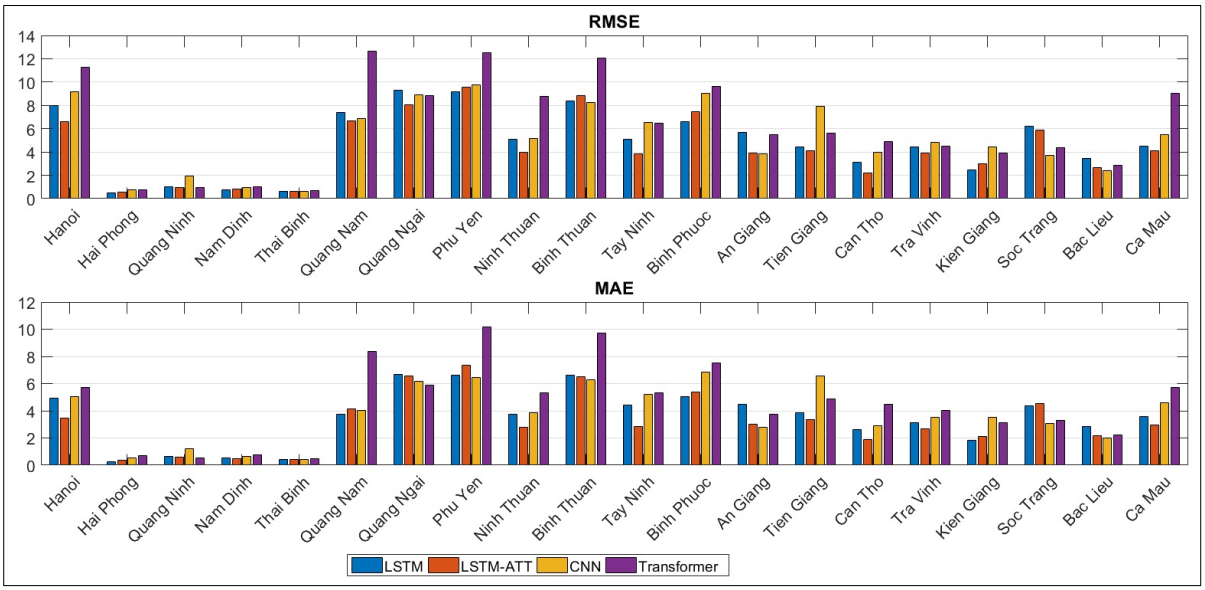

Figure 9: Error metrics for 1-month dengue fever predictions in 20 provinces in Vietnam. RMSE = Root Mean Squared Error, MAE = Mean Absolute Error, LSTM = Long Short-Term Memory, LSTM-ATT = attention mechanism-enhanced LSTM, CNN = Convolutional Neural Network.

All models were then ranked from 1 to 9 for each province, with lower numbers representing lower RMSE or MAE values. This allowed for the distributions of scores to be visualised in box and whisker plots, where mean rankings are shown as grey-outlined circles (Figure 10). Given the large number of evaluation provinces (20), this facilitated an easier comparison of the models. The LSTM-ATT model had an overall rank of first place, with mean place rankings of 1.60 for RMSE values and 1.95 for MAE values. The basic LSTM model came in second, with a mean RMSE-based ranking of 2.35 and a mean MAE-based ranking of 2.20, and the

CNN model came in third place on average, with a mean RMSE-based ranking of 3.10 and a mean MAE-based ranking of 2.70. In descending order, the subsequent models ranked as follows for both error metrics: transformer, XGBoost, Poisson regression, SARIMA, SVR, and SVR-L. As before, the attention mechanism appeared to improve the LSTM model performance.

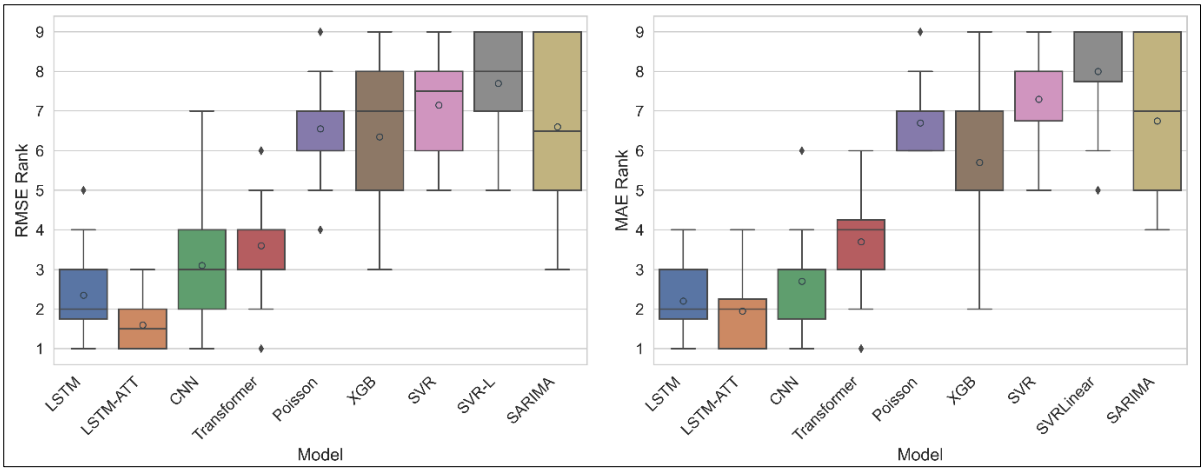

Figure 10: DF forecasting model rankings. Rankings are based on the relative scores for lowest RMSE or MAE in the prediction of dengue fever one month ahead. Box and whisker plots are shown, where grey-outlined dots indicate mean values. RMSE = root mean square error. MAE = mean absolute error. LSTM = long short-term memory. LSTM-ATT = attention mechanism-enhanced LSTM. CNN = convolutional neural network. Poisson = Poisson regression. XGB = Extreme Gradient Boosting. SVR = Support Vector Regressor with Radial Basis Kernel. SVR-L = Support Vector Regressor with Linear Kernel. SARIMA = Seasonal Autoregressive Integrated Moving Average.

Outbreak Detection at a One Month Lag

As the LSTM-ATT model had the best performance in previous results, its outbreak detection was assessed. LSTM-ATT detected all but two outbreak months (actual positives) in the test set—one in Thái Bình and one in Phú Yên (Figure 11, top). All outbreak months were detected

in Hà Nội, Quảng Nam, and Bình Phước, and there were only two false alarms raised—one in Hà Nội and one in Phú Yên. While there were no outbreaks in the other provinces to test the model on, LSTM-ATT also did not raise any false alarms. LSTM-ATT displayed high levels of precision, accuracy, sensitivity, and specificity in epidemic forecasting (Figure 11, bottom). There were high values for all performance metrics in the three provinces making up the majority of outbreak months—Hà Nội, Quảng Nam, and Bình Phước. There were, however, lower precision and sensitivity values of 0.5 in Phú Yên where one of the two outbreak months were missed, and 0 in Thái Bình where the one outbreak month was missed. The ability of the LSTM-ATT model to correctly assign non-outbreak months, as assessed by specificity and accuracy, was consistently high across all provinces with most values being 1.0. For many of the provinces, however, there were no outbreaks or predicted outbreaks, resulting in undefined values for precision and sensitivity.

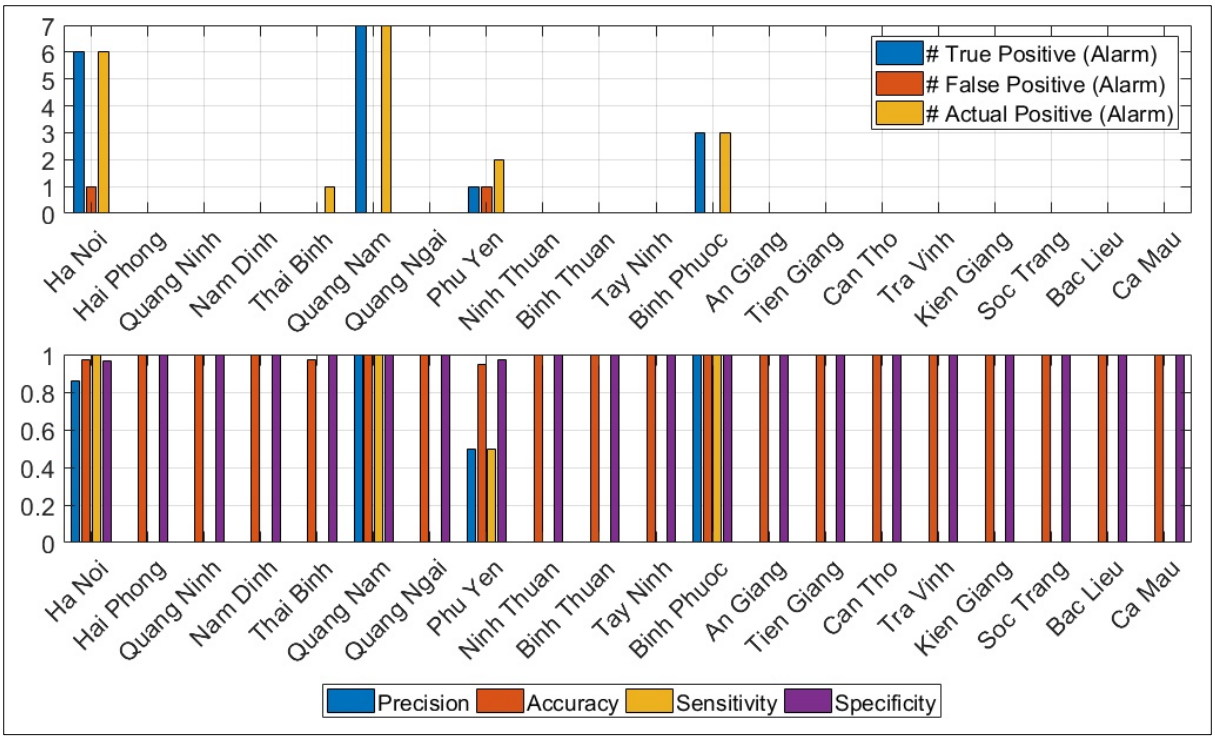

Figure 11: Dengue fever outbreak detection performance for the Attention Mechanism-enhanced Long Short-Term Memory model. Numbers of actual outbreaks, correct outbreak predictions (true positive) and incorrect outbreak predictions (false positive) for each province are shown (top). Additionally, prediction metrics (precision, accuracy, sensitivity, and specificity) for each province are displayed (bottom). If a province did not have any actual outbreaks in the evaluation period, the precision and sensitivity are not available.

3.2.2 Predicting Dengue Fever Multiple Months in Advance

Multi-step Forecasting

Following the methods used for predicting DF one month in advance, the performance of LSTM-ATT at forecasting DF two to three months ahead was tested (Figure 12). *k*-step notation is used, where *k* refers to the number of months in advance that the prediction is made. All *k*-step predictions were plotted for Hà Nội, Ninh Thuận, and Bình Phước to give a representative overview of the differences. In Hà Nội, the main visual difference between

predictions is in the outbreak peaking around month 21. The 2-month ahead prediction underestimated this the most, but the 3-month ahead prediction was slightly better. In the other two provinces there appears to be a progressive worsening as *k* increases. In Ninh Thuận, 2- and 3-month ahead predictions forecasted the month 23 outbreak peak to occur in later months, and in Bình Phước there is a clear false spike in cases at month 6 for the *k=3* prediction (Figure 12).

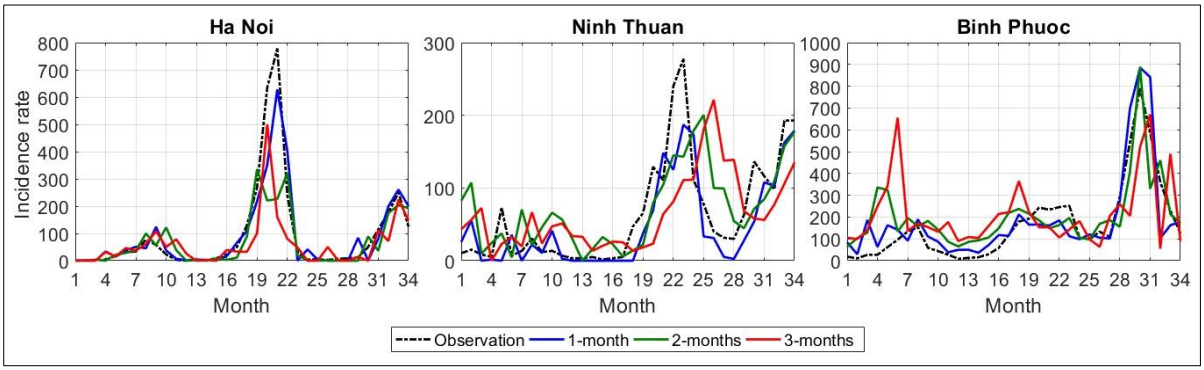

Figure 12: Multi-month ahead dengue fever predictions by the Attention Mechanism-enhanced Long Short-Term Memory model. Observation refers to the real rates of dengue fever incidence in the test set from 2014–2016. Predictions are plotted 1, 2, and 3 months in advance.

The error metrics mostly support the visual insights from Figure 12, with a slight decrease in RMSE between 2- and 3-month ahead predictions in Hà Nội and progressive increases from 1–3 month ahead predictions in Ninh Thuận and Bình Phước (Figure 13). In general, RMSE and MAE increased as *k* increased. For many provinces, however, the predictive performance was relatively consistent across 1–3 month ahead predictions (e.g., in Quảng Ngãi, Tây Ninh, and Bạc Liêu). Interestingly, in Bình Thuận, RMSE and MAE were lower for multi-step predictions.

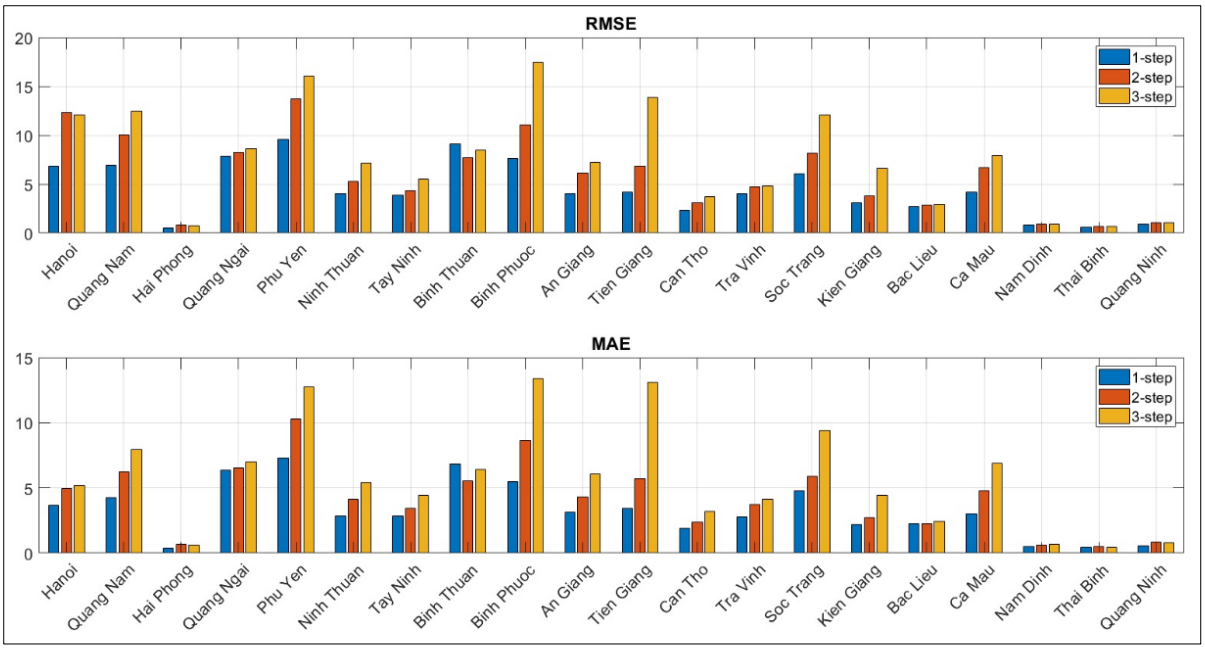

Figure 13: Multi-step prediction performance of the Attention mechanism-enhanced Long Short-Term Memory model for all provinces. RMSE and MAE values are provided as error metrics for all 20 provinces for predictions up to three months (steps) in advance. RMSE = root mean square error. MAE = mean absolute error.

Multi-step Outbreak Detection

As with one month ahead forecasting, outbreak detection was tested 2–3 months in advance. In line with the multi-step forecasting, the performance generally decreased as the number of months ahead increased (Figure 14). This was seen in Hà Nội, Quảng Nam, and Bình Phước. However, precision remained high at two months in these three provinces, meaning that the LSTM-ATT models missed very few outbreak months. Additionally, the specificity and accuracy in provinces with no outbreaks was consistently high, as it was for predictions one month ahead.

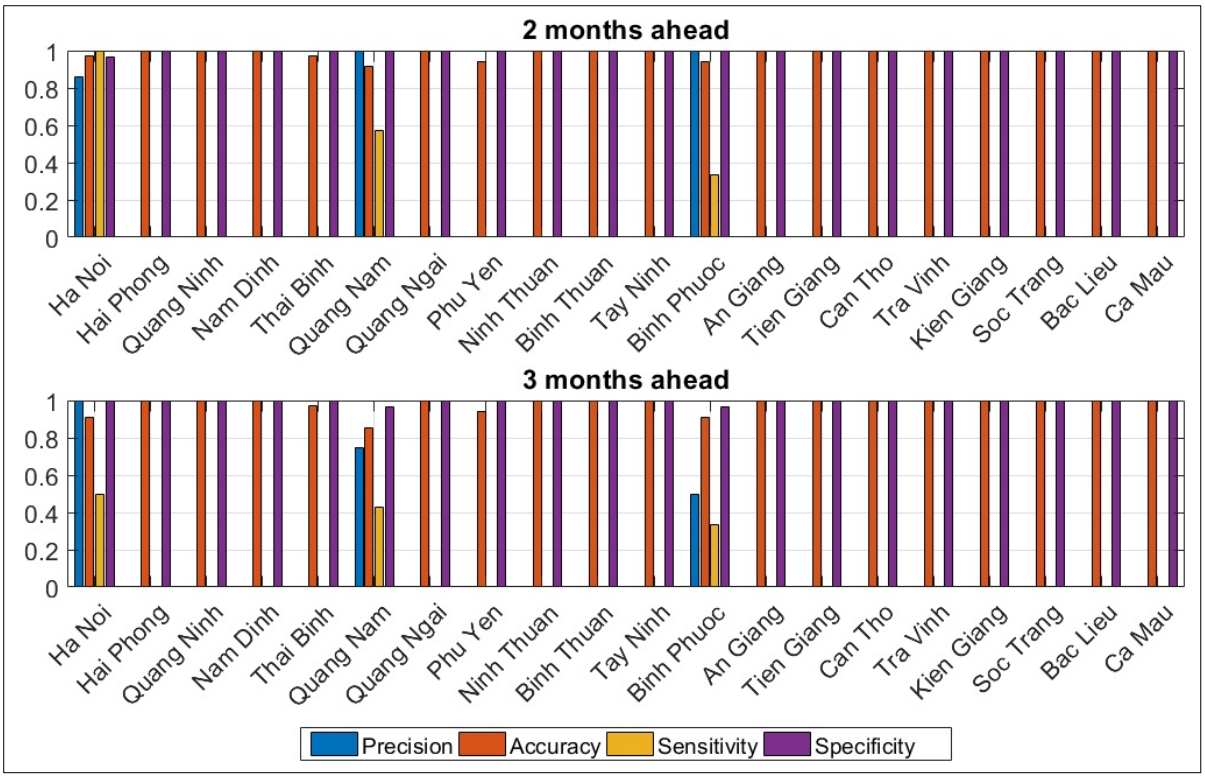

Figure 14: Multi-month dengue fever outbreak detection performance for the Attention Mechanism-enhanced Long Short-Term Memory model. Prediction metrics (precision, accuracy, sensitivity, and specificity) for each province are displayed. If a province did not have any actual outbreaks in the evaluation period, the precision and sensitivity are not available.

3.3 Diarrhoeal Disease

3.3.1 Hyperparameter Optimisation

SARIMA & SARIMAX

Hyperparameters were optimised by TPE to minimise RMSE for SARIMA models for each province, coming to near-optimum values early on in the hundred trials. Four of the six provinces had some trials with RMSE values above 60,000, however these were uncommon. Hyperparameter importance varied by location, though for three of the five provinces the

seasonal moving average component (Q) was the first or second most important. Differencing (d) and trend (t) were relatively unimportant for all provinces, however seasonal differencing was the most important hyperparameter in Kon Tum. The other parameters occasionally had high importance values on a province-by-province basis (Figure 15).

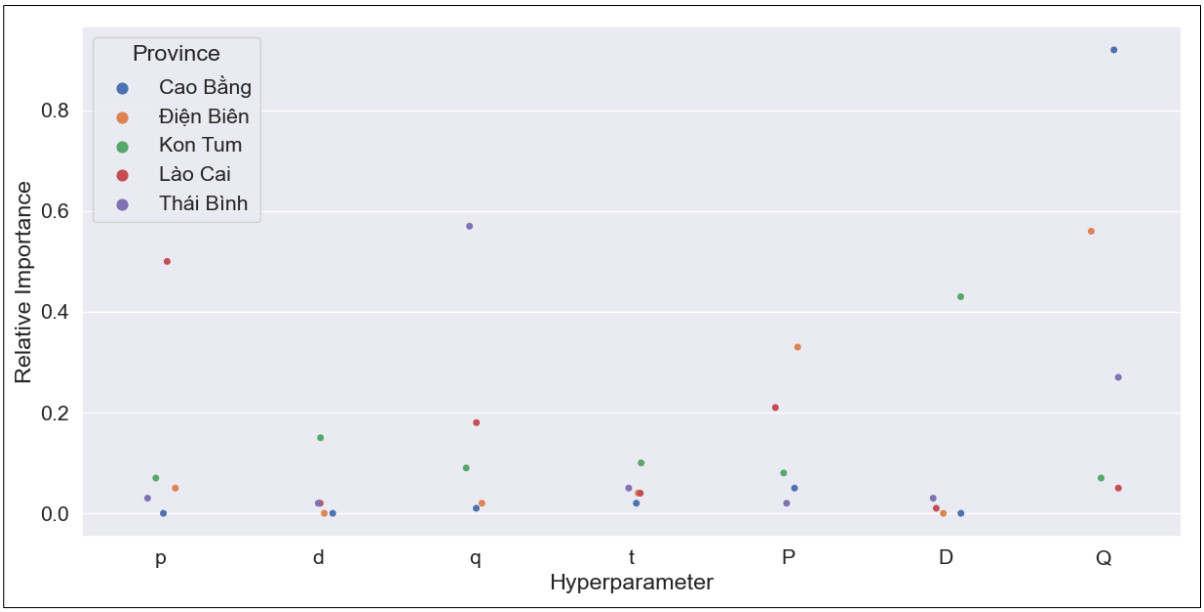

Figure 15: Relative hyperparameter importance values for SARIMA models. Importance values sum to one for each province, and indicate a hyperparameter's influence on minimising root mean square error.

Deep Learning Models

The TPE settled towards a near-optimum RMSE for LSTM, LSTM-ATT, and CNN models relatively early in the 100 trials. There were large variations in objective (RMSE) values for all models, however these were much greater for the LSTM and LSTM-ATT models than for the CNNs. These observations were consistent throughout the provinces, and the optimisation histories from Cao Bằng are provided as a supplementary example (Figure S1).

Learning rate was frequently a highly important hyperparameter in model optimisation, as was epochs, whereas hidden size and number of layers were commonly less important (Figure 16). As four hyperparameters were optimised for the CNNs as opposed to five for the other models, the relative values are not directly comparable. However, it can still be observed that the number of filters was relatively unimportant for RMSE minimisation in the CNN models compared to dropout rate. Additionally, lookback window appears to have been more important for the LSTM-ATT model than the standard LSTM model.

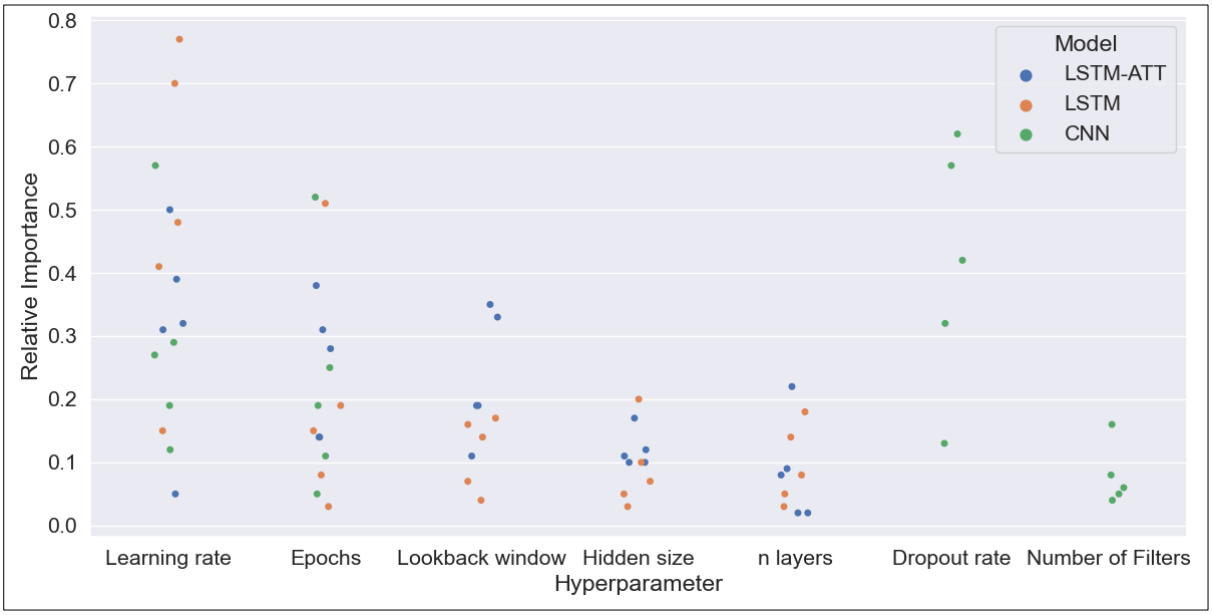

Figure 16: Relative hyperparameter importance values for deep learning models. Importance values sum to one for each province, and indicate a hyperparameter's influence on minimising root mean square error. Learning rate, epochs, lookback window, hidden size, and n_layers were optimised for Long Short-Term Memory (LSTM) and Attention mechanism-enhanced LSTM (LSTM-ATT) models for 5 provinces. Learning rate, epochs, dropout rate, and number of filters were optimised for Convolutional Neural Network (CNN) models for 5 provinces.

3.3.2 SARIMAX Associations

Total rainfall, sunshine hours, minimum absolute temperature, and influenza were found to be significantly associated with diarrhoea rates (Table 4). A 1mm increase in total rainfall was correlated with a mean increase in diarrhoea rates per 100,000 population of 0.324 one month later in Thái Bình. In the same province, a 1-hour increase in total sunshine preceded a mean 0.910-unit rise in diarrhoea rates per 100,000 population by one month. Minimum absolute temperature was associated with diarrhoea at a 1-month lag in Kon Tum, where an increase of 1°C corresponded to a rise of 18.3 diarrhoea cases per 100,000 population. Finally, a 1-unit increase in influenza rates per 100,000 population in Kon Tum was associated with a 0.0674-unit increase in diarrhoea rates per 100,000 population one month later, and a 0.148-unit increase three months later. In contrast, the same rise in influenza rates correlated with a 0.162-unit drop in diarrhoea rates per 100,000 population two months later in Điện Biên.

Table 4: Significant associations between climate factors and diarrhoea rates from the SARIMAX models. Full metrics are provided from the model results, which had all variables independently scaled from 0–1 using the Scikit-learn (version 0.24.2) MinMaxScaler (Pedregosa et al., 2011), in addition to the unscaled correlation coefficient. Significance thresholds were adjusted by Bonferroni adjustment for 6 comparisons per province (*: $p < 0.05$, **: $p < 0.01$). Min. Abs. Temperature = minimum absolute temperature., β_i = scaled correlation coefficient, Std. Error = standard error, z = z-score, CI = confidence interval, $\widehat{\beta_i}$ = unscaled correlation coefficient.

Province	Factor	All Factors Scaled to [0, 1]					Unscaled
		β_i	Std. Error	z	P>|z|	95% CI	$\widehat{\beta_i}$
Điện Biên	Influenza (Lag 2)	-0.533	0.111	-4.818	0.000**	(-0.75, -0.316)	-0.162
Thái Bình	Total Rainfall (Lag 1)	0.315	0.082	3.850	0.000**	(0.155, 0.476)	0.324
Thái Bình	Sunshine Hours (Lag 1)	0.268	0.088	3.041	0.002*	(0.095, 0.440)	0.910
Kon Tum	Min. Abs. Temperature (Lag 1)	0.558	0.090	6.212	0.000**	(0.382, 0.743)	18.3
Kon Tum	Influenza (Lag 1)	0.242	0.086	2.797	0.005*	(0.072, 0.411)	0.0674
Kon Tum	Influenza (Lag 3)	0.531	0.108	4.892	0.000**	(0.318, 0.743)	0.148

No other province-specific climate factors used in the SARIMAX models were found to have significant associations with diarrhoea at 1–3 month lags. As there were 1–3 month lags of two climate variables (or influenza rates) for each model, some of the significant associations presented here were not replicated in other provinces. For all models, influenza rates and measures of temperature were the most commonly selected predictors by random forest regression, with 3 uses each. Measures of rainfall and sunshine hours were each selected twice (Table S9).

3.3.3 Forecasting & Outbreak Detection

Forecasting One Month in Advance

In most provinces, the univariate SARIMA results appear to follow the observed diarrhoea rates better than the multivariate SARIMAX models (Figure 17). This is seen clearly in Điện Biên, where the multivariate model consistently underestimated diarrhoea rates, and in Kon Tum where there were large underestimates around months 0, 12, and 26. However, the SARIMAX model appears to perform slightly better in Lào Cai, where the addition of exogenous factors attenuates the false spike in cases near month 30. At month 5, there is a clear diarrhoea outbreak which is mostly missed by both models.

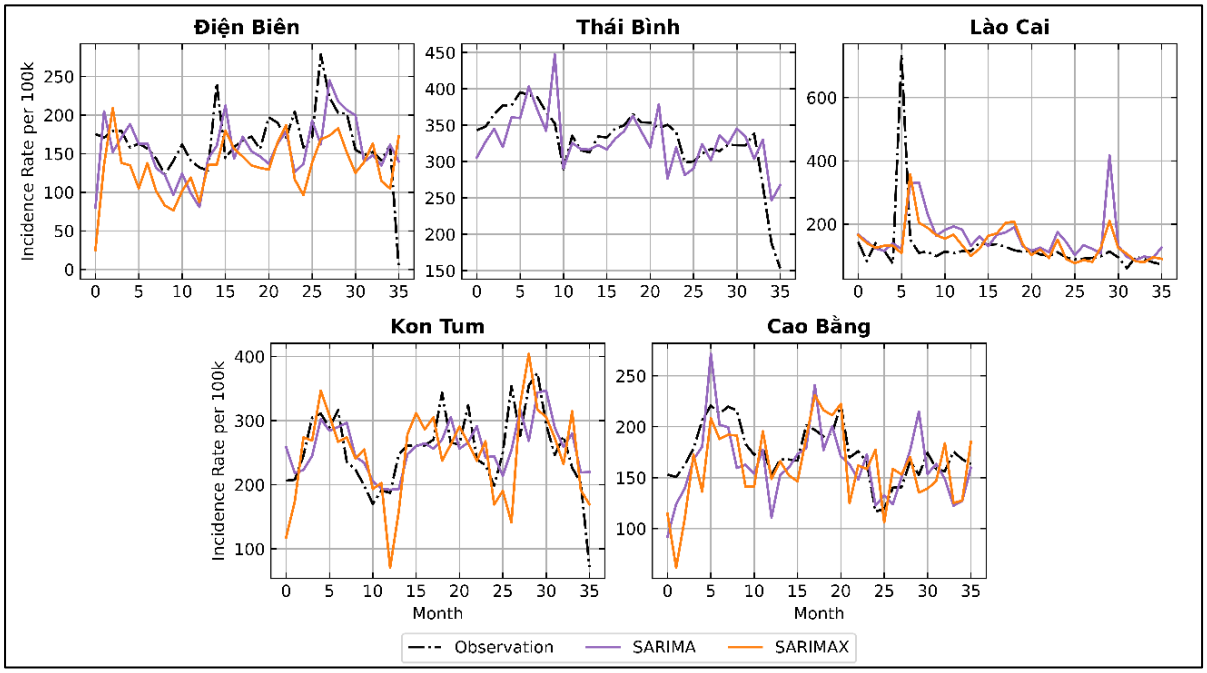

Figure 17: One-month ahead SARIMA and SARIMAX diarrhoea predictions for five provinces in Vietnam. Observation refers to the real rates of dengue fever incidence in the test set from 2014–2016. SARIMAX was not plotted for Thái Bình due to high inaccuracy obscuring the SARIMA results. SARIMA = Seasonal Autoregressive Integrated Moving Average. SARIMAX = SARIMA with exogenous regressors.

The error metrics confirm the higher performance of the SARIMA models, with the SARIMA models generally outperforming the SARIMAX models by RMSE, MAE, and MAPE metrics. This observation is reversed in Lào Cai, however, and the error scores are similar in Cao Bằng (Table 5). In Thái Bình, the SARIMAX RMSE was exceptionally high at 1081. Across the provinces, the SARIMA and SARIMAX performed best in Cao Bằng with MAPEs of 13.1% and 15.7%, respectively. The SARIMA model also performed well in Kon Tum, with an MAPE of 18.4%. The worst provinces were Điện Biên for SARIMA and Kon Tum for SARIMAX, where MAPEs of 284% and 6274% were observed. This, however, was impacted by the observed rate dropping to zero in Điện Biên at the last month of the test set.

Table 5: One-month ahead diarrhoea prediction performance metrics. Values are colour-coded for each province separately from the lowest value (darker green) to the median value (yellow) to the highest value (darker red). SARIMA = Seasonal Autoregressive Integrated Moving Average. SARIMAX = SARIMA with exogenous regressors. LSTM = long short-term memory. LSTM-ATT = attention mechanism-enhanced LSTM. CNN = convolutional neural network.

Province	Root Mean Square Error for Each Model				
	SARIMA	SARIMAX	LSTM	LSTM-ATT	CNN
Điện Biên	46.909	59.868	54.352	56.641	53.333
Thái Bình	37.273	1081.488	42.762	31.876	46.062
Lào Cai	131.215	118.172	114.680	109.585	112.034
Kon Tum	48.407	62.739	74.909	53.200	47.291
Cao Bằng	29.141	32.867	17.532	25.354	18.759
Province	**Mean Absolute Error for Each Model**				
	SARIMA	SARIMAX	LSTM	LSTM-ATT	CNN
Điện Biên	32.832	46.937	32.265	42.232	42.966
Thái Bình	26.831	734.728	33.563	23.717	33.953
Lào Cai	69.251	55.809	62.533	45.047	54.457
Kon Tum	37.338	48.358	61.846	39.222	36.908
Cao Bằng	22.948	26.285	14.207	20.163	13.649
Province	**Mean Absolute Percentage Error for Each Model (%)**				
	SARIMA	SARIMAX	LSTM	LSTM-ATT	CNN
Điện Biên	284.037	355.184	364.378	407.484	410.412
Thái Bình	9.420	218.909	11.452	8.434	12.504
Lào Cai	51.305	37.769	48.824	29.330	40.419
Kon Tum	18.391	6273.855	25.020	19.890	18.974
Cao Bằng	13.100	15.717	8.554	11.949	8.631

Results from the deep learning models were plotted together, separate from the SARIMA(X) models to avoid overplotting (Figure 18). In most of the plots, the deep learning models adhere closely to the line of observed diarrhoea rates. There are some visible differences between the deep learning models, such as a closer adherence by the LSTM model in Điện Biên, a large underestimation of rates by the LSTM model in Kon Tum, and an overshot of estimated rates around month 5 in Cao Bằng by the LSTM-ATT model. As with the SARIMA(X) models in Figure 17, the deep learning models mostly missed the large spike in diarrhoea rates at month 5 in Lào Cai.

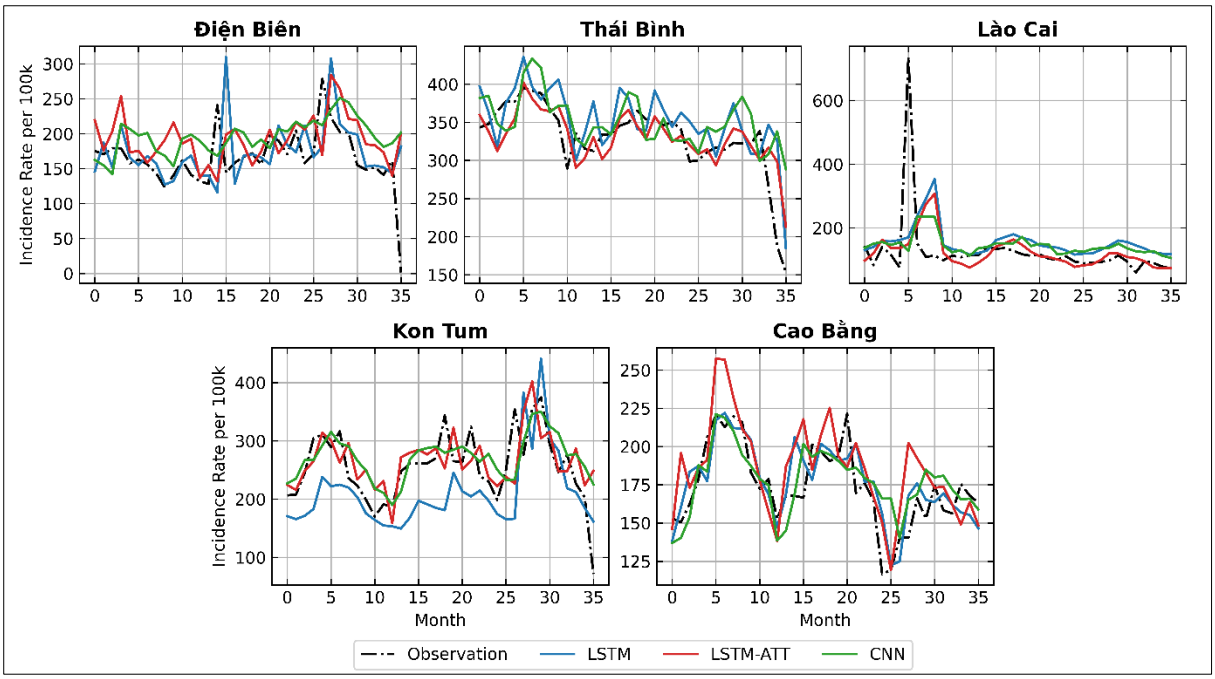

Figure 18: One-month ahead deep learning model diarrhoea predictions for five provinces in Vietnam. Observation refers to the real rates of dengue fever incidence in the test set from 2014–2016. LSTM = Long Short-Term Memory, LSTM-ATT = attention mechanism-enhanced LSTM, CNN = Convolutional Neural Network.

Considering all machine learning models together, LSTM-ATT displayed the highest performance in the most provinces, with the lowest RMSE, MAE, and MAPE values in both Thái Bình and Lào Cai (Figure 19). The CNN model had the lowest RMSE and MAE in Kon Tum, though the SARIMA model had a slightly lower MAPE. In Cao Bằng, LSTM had the lowest RMSE and MAPE while the CNN had a lower MAE. Despite this, the CNN model did show consistently strong forecasting performance. Across the entire selection of models, the province with the lowest relative error score was Thái Bình (LSTM-ATT MAPE: 8.43%), followed by Cao Bằng (LSTM MAPE: 8.55%), Lào Cai (SARIMA MAPE: 18.4%), Kon Tum (LSTM-ATT MAPE: 29.3%), and Điện Biên (SARIMA MAPE: 284%).

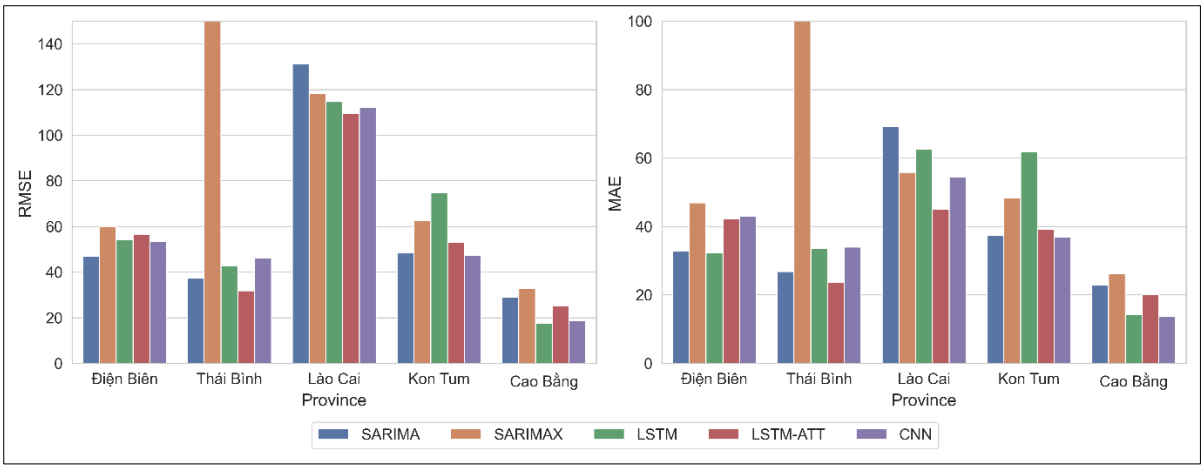

Figure 19: One-month ahead prediction performance metrics for all diarrhoea prediction models. RMSE and MAE values are provided as error metrics for predictions up to three months in advance. The plots were capped at 150 and 100, however the SARIMAX Thái Bình errors are greater than this limit. RMSE = root mean square error. MAE = mean absolute error. SARIMA = Seasonal Autoregressive Integrated Moving Average. SARIMAX = SARIMA with exogenous regressors. LSTM = long short-term memory. LSTM-ATT = attention mechanism-enhanced LSTM. CNN = convolutional neural network.

Subsequently, the models were ranked based on RMSE and MAE values (Figure 20). For RMSE-based scoring, CNN placed first with a mean ranking of 2.2, which was followed closely by LSTM-ATT with a mean ranking of 2.4. This was followed by SARIMA, LSTM, and SARIMAX with respective mean rankings of 2.8, 3.0, and 4.6. For MAE-based scoring, the situation was reversed—LSTM-ATT had the lowest mean ranking of 2.2 followed closely by CNN with 2.4. LSTM and SARIMA tied with mean rankings of 3.0, and SARIMAX came last again with a mean ranking of 4.4.

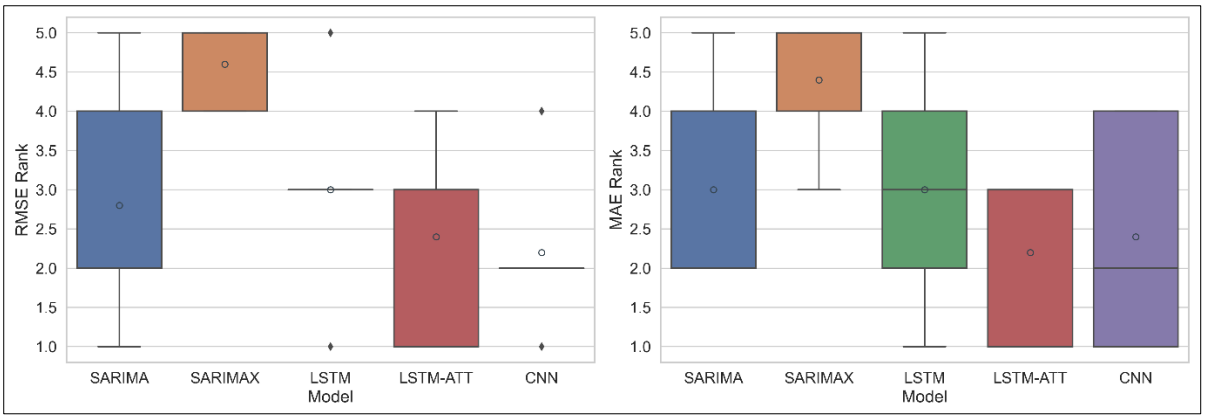

Figure 20: Diarrhoea forecasting model rankings. Rankings are based on the relative scores for lowest root mean square error in the prediction of dengue fever one month ahead. Box and whisker plots are shown, where grey-outlined dots indicate mean values. SARIMA = Seasonal Autoregressive Integrated Moving Average. SARIMAX = SARIMA with exogenous regressors. LSTM = long short-term memory. LSTM-ATT = attention mechanism-enhanced LSTM. CNN = convolutional neural network.

Outbreak Detection at a One Month Lag

Due to the LSTM-ATT model having the lowest MAE-based ranking, the lowest RMSE values in the most provinces compared to other models (2/5 provinces), and strong performance in DF forecasting, it was selected to move forward with outbreak detection. As with DF, outbreak months were defined as one standard deviation from the monthly mean of diarrhoea rates for a given province. There was only one outbreak month in the test set, which occurred in Lào Cai (Figure 21). The LSTM-ATT had a delayed forecast for this event, and predicted a false positive the following month. This resulted in precision and sensitivity scores of zero for Lào Cai. It also meant there were two incorrect predictions for true normal months, resulting in an accuracy of 0.944 and a specificity of 0.971. For the other four months, there were no outbreaks or predicted outbreaks for precision and sensitivity calculations, resulting in undefined values. LSTM-ATT correctly categorised all (non-outbreak) months for Điện Biên, Thái Bình, Kon Tum, and Cao Bằng, resulting in accuracy and specificity scores of 1.00.

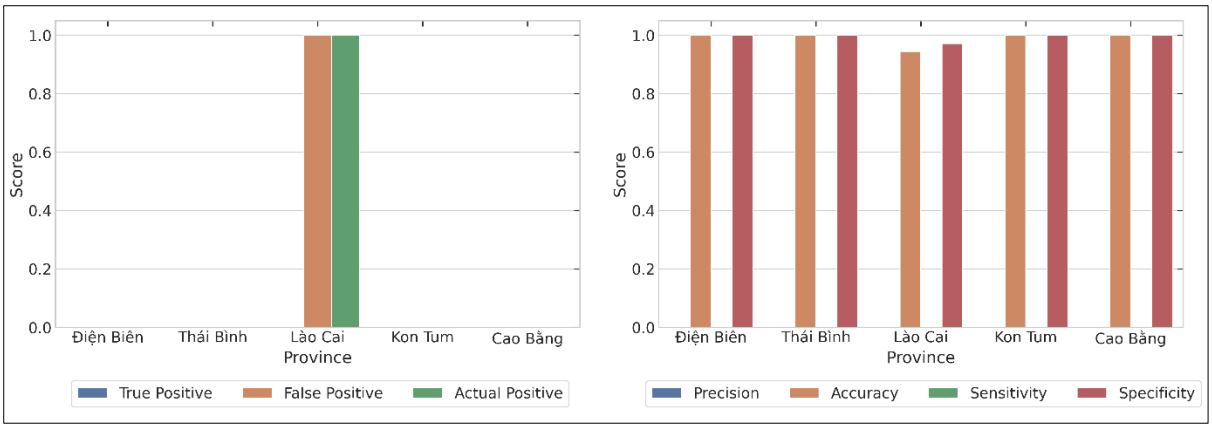

Figure 21: Diarrhoea outbreak detection performance for the Attention Mechanism-enhanced Long Short-Term Memory model. Numbers of actual outbreaks, correct outbreak predictions (true positive) and incorrect outbreak predictions (false positive) for each province are shown (left). Additionally, prediction metrics (precision, accuracy, sensitivity, and specificity) for each province are displayed (right). If a province did not have any actual outbreaks in the evaluation period, the precision and sensitivity are not available.

Multi-step Forecasting

In general, the forecasting worsened when projecting diarrhoea rates more months in advance. This was particularly noticeable in Kon Tum and Cao Bằng, where false spikes in predicted cases worsened as forecast lag increased. For example, this occurred around months 1 and 13 in Kon Tum, and around months 5–9 in Cao Bằng. However, the lines of predicted incidence are highly similar in Thái Bình and Lào Cai across all forecast lags (Figure 22).

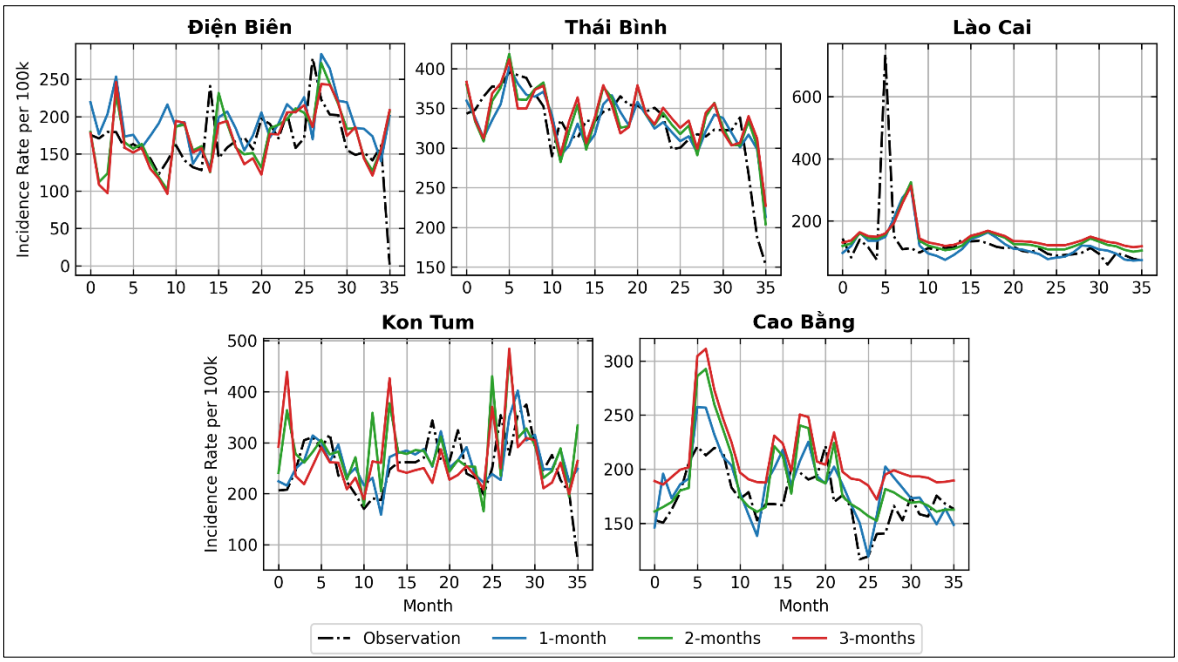

Figure 22: Multi-month ahead diarrhoea predictions by the Attention Mechanism-enhanced Long Short-Term Memory model. Observation refers to the real rates of dengue fever incidence in the test set from 2014–2016. Predictions are plotted 1, 2, and 3 months in advance.

The error metrics confirm a general worsening of predictive performance with increasing forecast lag, with some exceptions (Figure 23). These differences tended to be more pronounced in MAEs than RMSEs. RMSE increased as *k* increased for all provinces except Điện Biên, where 1-month predictions had the highest error. In Thái Bình, RMSE increased only slightly from 31.9 to 38.9, and MAPE from 8.43% to 10.8%, for 1–3 month ahead forecasts. In Lào Cai, RMSE performance was nearly identical across *k* values, ranging from 109.6 to 110.3, while MAPE had a greater increase from 29.3% to 42.4%. Furthermore, in Kon Tum 2- and 3-month ahead performance was similar (RSME: 85.6–85.7) but worse than 1-month projections (RMSE: 53.2). MAPEs ranged from 19.9–30.8%. Lastly, in Cao Bằng, 2-month ahead predictions only had a slight increase in error metrics from 1-month ahead predictions, with RMSE increasing from 25.4 to 31.2 and MAPE from 11.9% to 13.5%.

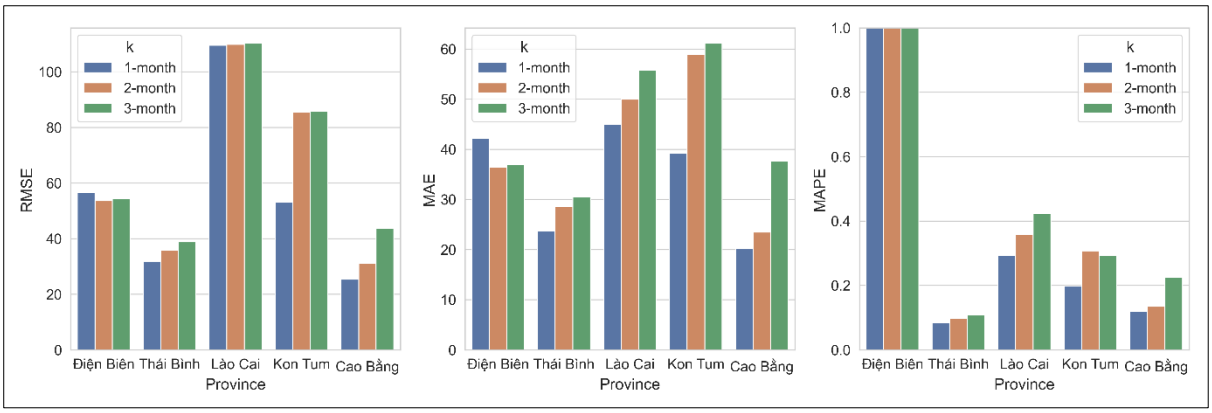

Figure 23: Multi-month ahead diarrhoea prediction performance metrics for LSTM-ATT. RMSE, MAE, and MAPE values are provided as error metrics for predictions up to three months in advance. The MAPE plot y-axis was capped at 1.0 (100%), though Điện Biên results exceed this. RMSE = root mean square error. MAE = mean absolute error. MAPE = mean absolute percentage error.

Multi-step Outbreak Detection

Accuracy and specificity remained high for the multi-step outbreak predictions (Figure 24). The only changes from 1–3 months ahead forecasts were in Kon Tum, where accuracy and specificity both dropped from 1.00 to 0.917 before rising to 0.972. As the true case numbers did not change, there was still only one outbreak month in Lào Cai and none in the other provinces resulting in undefined precision and sensitivity scores. Neither the 2- or 3-month ahead LSTM-ATT predictions correctly identified the outbreak in Lào Cai, resulting in precision and sensitivity scores of 0.

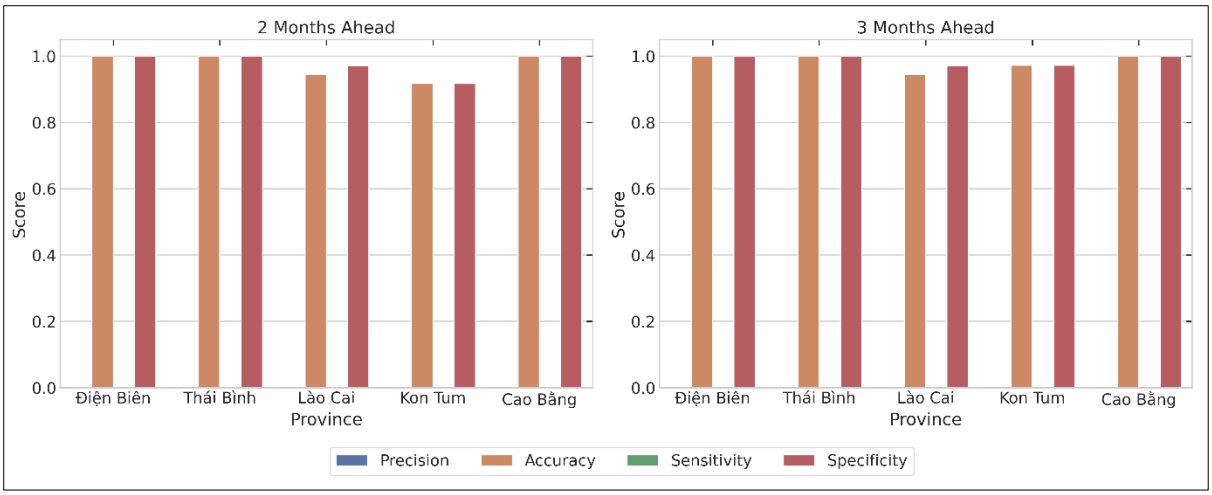

Figure 24: Multi-month diarrhoea outbreak detection performance for the Attention Mechanism-enhanced Long Short-Term Memory model. Prediction metrics (precision, accuracy, sensitivity, and specificity) for each province are displayed. If a province did not have any actual outbreaks in the evaluation period, the precision and sensitivity are not available.

4. Discussion

The first major section of this study examined the performance of various machine learning models on predicting DF incidence one to three month ahead in Vietnam, which was preceded by an analysis of the climate and disease datasets to provide context for the model results. Traditional and deep learning models were explored across 20 provinces for one-month ahead DF forecasts, where LSTM-ATT displayed the highest performance. Notably, the attention mechanism improved the results of the basic LSTM model in most provinces. As the results suggested it to be model with the strongest overall performance, LSTM-ATT was tested for one-month outbreak detection, as well as for 2–3 month ahead prediction and outbreak detection.

The second section of this study aimed to deliver an improved disease forecasting model, focusing on the prediction of diarrhoeal disease. LSTM-ATT, LSTM, and CNN were used due to their strong performance in section one. Additionally, SARIMA was included because of potential for improvement through optimisation of its many hyperparameters and the addition of exogenous regressors. TPE modelling from the Optuna Python library (Akiba et al., 2019) was implemented for hyperparameter optimisation as an improvement upon trial-and-error selection used for the DF models. SARIMA and SARIMAX models were included to check whether TPE optimisation could develop models with performance competitive to the deep learning models. As before, forecasting and outbreak detection were examined at 1–3 month lags.

Models were optimised on using RMSE, though evaluation metrics are also provided for MAE and MAPE. RMSE is potentially a more useful metric than MAE, as it attributes greater weights to larger errors. For disease forecasting, missing one large outbreak is arguably much worse than many having small inaccuracies such as predicting 62 cases per 100,000 instead of 67. However, MAE is a more interpretable metric, so it was included to observe on average how many cases per 100,000 a prediction was off by. Lastly, MAPE was included as a relative metric to allow comparison between different provinces and with other studies. MAPE, however, has disadvantages when observed rates are small or zero. MAPE increases towards infinity as observed rates approach zero, and this can inflate MAPEs and falsely suggest poor model performance in the context of disease forecasting. This was seen in the DF predictions, where observed rates of zero and close to zero occurred regularly.

4.1 DF Forecasting

4.1.1 Climate Factors as Predictors of DF

Mosquito development and activity are affected by climate factors, which is part of the reason for the associations between climate factors and DF incidence. Recursive feature elimination selected two climate factors as predictors for each province, resulting in all variables except minimum relative humidity being used in at least one province. Therefore, measures of each weather descriptor—rainfall (16 uses), temperature (15 uses), sunshine hours (4 uses), evaporation (3 uses), and humidity (2 uses)—were important for prediction. Firstly, rainfall has previously been identified as a risk factor for DF outbreaks, as excess water can collect in various empty containers and create sites for mosquitos to breed in (Tuyet-Hanh et al., 2018a). In contrast, very heavy rain may wash out breeding grounds (Wang et al., 2014). Moreover, Barrera et al. (2011) found mosquito density to correlate with both high rainfall and DF incidence. Secondly, *Ae. aegypti* mosquitos have been shown to have increased rates of development as temperatures rise towards 30°C. However, very high temperatures appear detrimental, as development rates drop after 40°C (Eisen et al., 2014). Mosquitos may also bite more frequently with rising temperatures (Scott et al., 2000). When looking at the overall effects, dengue epidemic potential is predicted to rise as average temperature rises up to 29°C, though this temperature may decrease with higher diurnal temperature ranges. Additionally, in tropical regions where temperatures are near this threshold (e.g., Vietnam), lower diurnal temperature ranges result in a higher epidemic potential than high ranges (Liu-Helmersson et al., 2014). Lastly, while fewer studies have investigated the effects regarding sunshine hours and evaporation, high humidity levels have been shown to favour mosquito survival (Lega et al., 2017). These associations could explain, in part, the usefulness of the climate factors from

our research in forecasting DF. Moreover, the deep learning models may have performed well because of their ability to process the complex, non-linear relationships.

The length of lookback window used is supported by previous research into the time lags between altered weather conditions and DF incidence. A period of three months was used, which had lower error scores than other lengths from one to eighteen months. Delays between climate factors and DF incidence have been described as 0–3 months for rainfall (Do et al., 2014; Lee et al., 2017b; Pham et al., 2011; Phung et al., 2015c; Phuong et al., 2016; Wang et al., 2014), 0–2.5 months for temperature (Colón-González et al., 2011; Do et al., 2014; Lee et al., 2017b; Lowe et al., 2018; Pham et al., 2011; Phung et al., 2015c; Phuong et al., 2016; Wang et al., 2014), 0 months for evaporation (Tuyet-Hanh et al., 2018b), and 0–3 months for humidity (Pham et al., 2011; Phung et al., 2015c; Phuong et al., 2016; Tuyet-Hanh et al., 2018b, p.; Wang et al., 2014). A three month period, therefore, appears to capture all relevant delayed effects caused by incubation periods, altered mosquito development and behaviour, or altered human behaviour.

4.1.2 Forecasting DF Rates One Month in Advance

LSTM-ATT displayed the highest performance in DF forecasting one month ahead, followed by LSTM and CNN in second and third place, respectively. The traditional statistical and machine learning models—Poisson Regression, SVR, SVR-L, XGBoost, and SARIMA—had higher error metrics than the four deep learning models in most provinces, with a small number of exceptions. There are many factors at play in infectious disease forecasting, including

extraction of the most relevant predictors; the complex, non-linear impact of climate factors on mosquitoes; the impact of climate factors on human behaviour; and the extrinsic and intrinsic incubation periods of the dengue virus. Deep learning models provide the advantages of extracting the most relevant predictors, remembering information across many time steps, and deciphering complex data patterns with minimal manual feature engineering necessary (Bengio et al., 2013; Schmidhuber, 2015). This could explain the higher performance observed for the deep learning models.

In addition to predicting disease rates, the ability to predict outbreaks has the potential to be highly valuable for DF management. Therefore, as the LSTM-ATT model had the best performance in previous results, its outbreak detection was assessed. Based on methods from previous forecasting studies which defined outbreak thresholds as rates *n* standard deviations above the monthly mean of cases (Brady et al., 2015; Cheng et al., 2020), we set *n=1*. This allowed for the detection of smaller-scale outbreaks instead of only major ones. Incidence rates of DF vary greatly between provinces in Vietnam, so province-specific thresholds were chosen over single or fixed thresholds which have been implemented in other works (Cheng et al., 2020; Hii et al., 2012).

LSTM-ATT was able to detect almost all DF outbreak months, with the exception of one in Thái Bình and one in Phú Yên. Similarly, it correctly identified the vast majority of normal months, with only two false alarms raised. More outbreaks were observed in the central provinces than other regions, which was unanticipated—more were expected in the South due to hotter, sunnier conditions at low altitude. This could be partially due to local weather conditions favouring DF transmission, such as those in Quảng Nam which had the highest

number of outbreak months. From examining median climate metrics during the test set, Quảng Nam had the 2nd highest maximum daily rainfall, 5th highest number of rainy days, and 12th highest monthly rainfall out of the 55 provinces for which climate data was available. Moreover, the high temperatures in southern provinces may have been too high for optimal DF transmission (Eisen et al., 2014; Liu-Helmersson et al., 2014; Scott et al., 2000). Additionally, other factors may have affected the number of outbreaks. These could include higher availability of empty containers for water to pool in for mosquito breeding, lower use of mosquito nets and repellents, or different distributions of dengue virus serotypes causing more severe (and therefore registered) cases.

Our findings showing the high performance of LSTM models in climate-based DF forecasting are supported by previous works. A study by Pham et al. (2018) in Kuala Lumpur, Malaysia, found an LSTM using a genetic algorithm for hyperparameter optimisation to outperform linear regression and decision tree models. More recently, Xu et al. (2020) compared LSTM to SVR, GAM, GBM, and BPNN models, finding LSTM to outperform the other techniques and transfer learning to decrease error rates in areas with low incidence of DF. Similarly, our LSTM models performed strongly compared to the other DF prediction models. The improved metrics achieved by the LSTM-ATT model suggest that attention mechanisms could improve LSTM models in other disease forecasting contexts.

Finally, the Transformer model generally had much higher RMSE and MAE values than the other deep learning models. This was unanticipated given the strong performance compared to LSTM models in other contexts (Zeyer et al., 2019). Transformer models use self-attention mechanisms, which allows parallel processing of the data without it being in order. The poor

performance observed in this study could be due to suboptimal model development. Alternatively, the processing of out of order data could be negatively impacted by the seasonality of the data.

4.1.3 Multi-step Forecasting

LSTM-ATT was used for multi-step forecasting due to outperforming the other models in one-step predictions. In general, the performance of LSTM-ATT decreased when forecasting incidence and outbreaks further in advance. There were some exceptions, such as in Bình Thuận where error rates decreased. A probabilistic superensemble of generalised linear mixed models developed by Colón-González et al. (2021) was able to forecast prospective DF case numbers up to six months in advance throughout Vietnam, with an average continuous rank probability score of 110 outperforming baseline and individual models at lead times of 1–3 months. Outbreak detection was also assessed, with average accuracy and sensitivity scores of 73% and 68% for outbreaks more than two standard deviations above the mean. Due to the different outbreak threshold and the results being averaged across 1–6 month lags, the results are not directly comparable with the results presented here. However, the cost-loss analysis presented in the study suggests that the superensemble model was accurate enough to provide relative value over using no forecast to mitigate DF outbreaks in the majority of provinces. Future work to benchmark the superensemble model against the models presented here may prove constructive, especially considering we have not come across any other long-term DF prediction models in Vietnam.

A few previous studies in Singapore have shown results from long-term DF forecasting models, though they are limited in number. Hii et al. (2012) used a Poisson multivariate regression model to predict DF outbreaks four months ahead in Singapore at a national level. The Receiver Operating Characteristics (ROC) area under the curve (AUC) was high at 0.98. However, the test period was only one year long and contained only one outbreak, which reduces the robustness of the assessment. A more recent national-scale study by Shi et al. (2016) predicted weekly DF incidence up to three months ahead using LASSO regression models. The models used climate factors in addition to mosquito surveillance data for prediction, resulting in MAPE values of 17–24% for forecasts 1–3 months in advance. Chen et al. (2018) also used LASSO regression, but at a residential level in contrast with the two previous studies. Spatiotemporal case data, building age, meteorological conditions, and Normalised Difference Vegetation Index were used to construct the model, and AUC values ranged from 0.88 to 0.76 for 1–12 week forecasts. Due to the different reporting metrics, most of these studies are not directly comparable with our results. The MAPE values reported by Shi et al. (2016) were much lower, however they were reported at a national level. Therefore, it was less likely for the models to encounter months with zero observed cases of DF, inflating the MAPEs as occurred in our models.

4.2 Diarrhoeal Disease

4.2.1 Hyperparameter Optimisation

The Optuna TPE found the seasonal moving average component (Q) to regularly be the most important hyperparameter for SARIMA optimisation. Forecast errors for the same month in

previous years may be valuable information because of the annual seasonality of diarrhoea incidence. Trend and differencing were observed to be relatively unimportant, which was somewhat anticipated. No trend was observed visually or by augmented Dickey-Fuller test in Thái Bình (p = 0.008), Lào Cai (p = 0.002), Kon Tum (p = 0.007), or Cao Bằng (p = 0.013), and the trend in Điện Biên was eliminated after one differencing (p = 4.36e-15). Therefore, only ranges of 0–1 were optimised for *d*, leaving little room for strong alterations to the time sequences compared to the larger ranges for other parameters. The optimised values matched those suggested by the augmented Dickey-Fuller test.

Learning rate, and epochs to a lesser extent, were regularly highly important hyperparameters in tuning the deep learning models. Notably, lookback window was more important for LSTM-ATT than LSTM. Adding the attention mechanism was intended to attenuate the loss of information between previous months, so the difference in lookback window suggests this could have occurred. Other studies examining hyperparameter importance have found some similar results with regards to learning rate. Bergstra and Bengio (2012) assessed relative hyperparameter importance for random search-based hyperparameter optimisation on neural network experiments on nine image classification tasks based on variations on MNIST, rectangles, and convex image datasets. Relative importance changed based on the dataset, but learning rate remained highly important throughout the tasks. Hidden units had low relevance for most tasks, but was the second most important hyperparameter for the rectangles task. In another study, learning rate was identified as the most important hyperparameter for the lenet supervised learning algorithm and the deep-autoencoder unsupervised model on the CIFAR10 dataset (Jia et al., 2016). While these experiments were on image classification and not time series tasks, they appear relatively in line with the findings from the diarrhoea model optimisations.

4.2.2 Associations between Diarrhoea, Climate Factors, and Influenza

Minimum Absolute Temperature

Significant associations were found between several lagged climate factors and diarrhoea rates, including positive correlations with minimum absolute temperature. A 1°C increase in minimum absolute temperature was associated with a mean increase of 18.3 diarrhoea cases per 100,000 population one month later in Kon Tum ($p < 0.001$). This complements other studies which have found positive correlations between increased average temperature and diarrhoea 0–2 months ahead in Vietnam (Phung et al., 2018, 2015c) and in Kon Tum specifically (Lee et al., 2017a). While inverse and inverted-v correlations with average temperature have been described in Australia (D'souza et al., 2008) and Japan (Onozuka and Hashizume, 2011), respectively, this may be due to differences in geography, climate, and sanitation infrastructure.

A meta-analysis of 26 studies by Carlton et al. (2016) found an overall positive correlation between ambient temperature and diarrhoea or bacterial diarrhoea, but not viral diarrhoea. This may help to explain the relatively small effect of minimum absolute temperature on diarrhoea rates, as viral pathogens represent the main burden of diarrhoeal disease in Vietnam. Bacterial diarrhoea, by extension, is less common (Anders et al., 2015; Nguyen et al., 2004; Thompson et al., 2015b). The survival of diarrhoeagenic pathogens such as rotavirus, *E.coli*, and *Salmonella* species in surface and ground waters has been shown to decrease with rising temperatures (Blaustein et al., 2013; El-Senousy et al., 2014), suggesting an alternative reason for the association. However, other research has found increased diarrhoeagenic *E. coli* contamination of food at high ambient temperatures in Bangladesh (Parvez et al., 2017; Black et al., 1982). Alternative explanations include increased water usage, which may lead to

increased exposure to pathogens, and worse hygiene behaviour in hotter weather (Checkley et al., 2000).

Total Rainfall

Total rainfall and number of sunshine hours at a 1-month lag were found to positively correlate with diarrhoea rates in Thái Bình. A 1mm increase in total rainfall was significantly associated with a 0.324 unit increase in diarrhoea incidence per 100,000 population in the next month ($p < 0.001$). Given that the 2016 population of Thái Bình was 1.79 million people, a not-uncommon monthly increase in rainfall of 100mm would correlate with an increase of 580 cases in the final year of the test set. The SARIMAX model that identified this association performed very poorly, which should be acknowledged in the context of these results. Nevertheless, the association is worth exploring. Previous research has also found increased total monthly rainfall to precede increased diarrhoea rates by 1 month (Phung et al., 2015c) or other lags up to 2 months (Phung et al., 2017; Wangdi and Clements, 2017). The percentage of households in Vietnam using hygienic latrines rose from 55% in 2002 to 94% in 2020, with rural rates rising from 44% to 91% in the same period. Similarly, the percentage of households with access to clean water increased from 78% in 2002 to 97% in 2020, with slightly lower figures of 74% to 96% in rural areas (General Statistics Office of Vietnam, 2021b). Increased rainfall may contaminate water sources for households without access to clean water, and low latrine use may contribute to such contamination and diarrhoea rates. Improving water hygiene access in the country, therefore, could potentially mean a weakening association between rainfall and diarrhoea rates. However, even with hygienic water infrastructure in place, high rainfall and associated flooding can overwhelm sewage systems and cause overflow into rivers and other bodies of water (Ding et al., 2013).

Another possible factor influencing diarrhoea rates is the impact of rainfall on human behaviour. Rainy periods could cause people to stay indoors in close-contact, facilitating the transmission of diarrhoea-causing illnesses such as rotavirus infection. Heavy rainfall and flooding have been shown to reduce access to healthcare and medication in Hà Nội, caused by damage to transport infrastructure, damage to healthcare infrastructure, or insufficient money. This was accompanied by increased rates of communicable diseases and greater increases in reported hypertension in flooded compared to non-flooded areas (Bich et al., 2011). In coastal Vietnam, tropical storms have previously caused significant damage to agriculture crops, irrigation systems, and schools, as well as negative effects on mental health (Nguyen et al., 2017). Therefore, infrastructure damage from heavy rainfall and flooding could possibly leave households vulnerable to diarrhoeal disease because of reduced access to healthcare, income and food insecurity from crop damage, and associated detriment to mental and physical health.

Sunshine Hours

Sunshine hours were also positively associated with diarrhoea rates in Thái Bình, with a rise in cases of 0.910 per 100,000 population per 1-hour increase in monthly sunshine hours ($p < 0.01$). Relatively fewer studies have been examined this relationship. In China, sunshine hours above 150 per month correlated with higher risk of diarrhoea in those over 20 years at a 0–1 month lag (Fang et al., 2019). Similarly, Islam et al. (2009) found sunshine hours and temperature to synergistically correlate with increased cholera incidence in the same month, with high levels of one compensating for low levels of the other. However, it is unclear if this pattern also applies to diarrhoea associated with other pathogens. Findings of a negative correlation between sunshine hours and bacillary dysentery in the same month suggest it may not (Zhao et al., 2016). In a study by Oh et al. (2021), sunshine rate was associated with diarrhoea caused by *Clostridioides difficile* toxin B, but not *E. coli* O157:H7, *Campylobacter*

spp., or *Clostridium perfringens*, for the same month. The mechanistic reasons for the association between sunshine hours and diarrhoea found here are likewise unclear, though it could be due to a positive correlation between sunshine hours and temperature. Alternatively, as mentioned previously, the poor SARIMAX performance in Thái Bình may suggest this association to be a false positive.

Influenza Rates

A 1-unit increase in influenza rates per 100,000 population significantly correlated with a 0.0674-unit increase in diarrhoea rates per 100,000 population at a 1-month lag ($p < 0.01$) and a 0.148-unit increase at a 3-month lag in Kon Tum ($p < 0.001$). In contrast, a negative correlation was observed in Điện Biên, with a 1-unit increase in influenza rates preceding a -0.162-unit decrease in diarrhoea rates at a 2-month lag ($p < 0.001$). We have not come across other diarrhoea forecasting papers using influenza rates as a predictive factor. While diarrhoea can be a symptom of influenza, influenza is generally not attributed as a major cause of the diarrhoeal burden in Vietnam or elsewhere (Anders et al., 2015; Huyen et al., 2018; Isenbarger et al., 2001; Kotloff et al., 2013; Nguyen et al., 2004; Thompson et al., 2015b). This suggests another reason for the correlations found here. Gilca et al. (2012) designed a multivariable Box-Jenkins transfer function model to explore associations between influenza diagnoses and diarrhoea associated with *Clostridium difficile* infections. Positive influenza tests were found to be positively correlated with *C. difficile*-associated diarrhoea rates one and twelve months later, independent of antibiotic prescription. A positive feedback loop has previously been described where diarrhoea inflicts malnutrition upon children, thereby making them more likely to suffer from further incidence of infectious diseases, malnutrition, and diarrhoea (Troeger et al., 2018). Another possible explanation is diarrhoea and influenza being mutually impacted by a common external factor. Both diarrhoea and influenza rates have been found to

correlate with flooding, with influenza rates rising in the first ten days after flooding and infectious diarrhoea rising 10–60 days after flooding depending on pathogenic cause (Ding et al., 2019). This is in line with the positive association between influenza rates and diarrhoea at 1- and 3-month lags in Kon Tum.

The negative association at 2 months in Điện Biên contrasts the findings in Kon Tum and may represent either uncertainty in these correlations, or provincial differences in the impact of influenza. Điện Biên is a mountainous province in the Northwest region, while Kon Tum is located in the central highlands. In addition to the significant meteorological differences we observed between central and northern Vietnam for temperature, humidity, evaporation, and sunshine hours, the pathogenic causes of diarrhoea can vary between regions in Vietnam. For example, Kon Tum has historically had the highest rates of bacillary dysentery (Lee et al., 2017a), and slight differences in rotavirus-positive diarrhoea have been observed between northern, central, and southern Vietnam (Huyen et al., 2018). Therefore, the differences in influenza associations could potentially be related to the epidemiological differences in diarrhoea transmission.

4.2.3 Forecasting Diarrhoea Rates One Month in Advance

As with the DF predictions, LSTM-ATT was found to be the strongest prediction model for 1-month ahead forecasts based on MAE-based rankings. Additionally, the attention mechanism improved results over the simple LSTM model in three out of the five provinces. Five is a relatively small number of provinces to differentiate the model performances, given that their

performance was competitive in many cases. The CNN model had consistently strong performance too with the best mean RMSE-based ranking, which mirrors findings from a previous study. Abdullahi and Nitschke (2021) showed CNN models to result in higher accuracy than LSTM models when forecasting daily diarrhoea incidence in South Africa. There seems to be few other studies for comparison which have ranked deep learning models in diarrhoea forecasting. However, other studies have mirrored our findings of self-attention improving the performance of LSTM models. Zhu et al. (2019) developed an attention mechanism-enhanced multichannel LSTM model for predicting influenza rates from climate factors, which outperformed multichannel LSTM, regular LSTM, and traditional RNN models.

While they had higher error metrics in most provinces, the TPE-optimised traditional machine learning models—SARIMA and SARIMAX—were occasionally able to compete closely with the deep learning models. This was observed when SARIMA outperformed all other models as in Điện Biên, and also from RMSEs close to the lowest value in Cao Bằng (SARIMA: 48.4 vs LSTM: 47.3) and Lào Cai (SARIMAX: 118 vs LSTM-ATT: 110). Surprisingly, the multivariate models performed worse in most provinces, which may be due to nonlinear relationships between climate factors and diarrhoea incidence that could not be modelled by the linear regressions. Our findings seem similar to other diarrhoea and time-series forecasting studies, which have found ARIMA(X) and SARIMA(X) models to occasionally compete with deep learning models but to have inferior overall performance metrics. For example, in China Fang et al. (2020) found a random forest model using climate factors to outcompete ARIMAX in diarrhoea prediction (21% vs 28% MAPE). Similarly, Jia et al. (2019) showed a LSTM model using seasonal and morbidity data to outperform an ARIMA model, in addition to linear regression and XGBoost models, in the prediction of ten infectious diseases in China including diarrhoea and dysentery. More recently, Kırbaş et al. (2020) found LSTM to be more accurate

than nonlinear autoregression neural network and ARIMA models in predicting 7-step coronavirus disease 2019 (COVID-19) cases in Belgium, Germany, France, Denmark, Switzerland, the United Kingdom, Finland, and Turkey. Moreover, in a paper on building energy load forecasting, ARIMAX had higher performance in one multi-step forecasting test, but overall a CNN model increased relative accuracy by 22.6% (Cai et al., 2019). It seems, therefore, that deep learning models regularly outperform ARIMA-based models in time-series forecasting for diarrhoea, for other infectious diseases, and in other contexts.

While the performance of the diarrhoea models varied throughout the provinces and between models, MAPEs of as low as 8.43% were obtained in Thái Bình (by LSTM-ATT). The lowest relative scores for the best model (determined by RMSE) in the other provinces were 8.55% in Cao Bằng by LSTM, 19.0% in Kon Tum by CNN, 29.3% in Lào Cai by LSTM-ATT, and 284% in Điện Biên by SARIMA. The LSTM-ATT model failed to identify the one outbreak in Lào Cai, and there were no other outbreaks to evaluate on. It identified non-outbreak months correctly in most cases, however, with the lowest accuracy and specificity scores being 0.944 and 0.971. The lack of outbreaks may indicate that outbreak detection is less suitable for diarrhoea models than DF models, as there appears to be a more constant presence of diarrhoea rates with less specific outbreaks in the period of 1997 to 2016 in Vietnam (Figures 6, 7). This aside, the models appear to be suitably high performing for useful forecasting tools in some provinces in Vietnam (e.g., Thái Bình and Cao Bằng), but may be of limited utility in others without further development.

4.2.4 Multi-step Forecasting

MAPE values of as low as 10.8% were obtained for three-month ahead predictions, as observed in Thái Bình, indicating LSTM-ATT to be a strong model for long-term diarrhoea forecasting in some provinces. MAPE remained low for predictions two months ahead in Cao Bằng, as well, rising from 11.9% to 13.5%, before increasing to a more moderate value of 22.5% for predictions three months in advance. Forecasts two and three months ahead were similar in Kon Tum, with values of 30.8% and 29.3%, respectively. These are approximately a 50% increase above the 1-month prediction MAPE of 19.9%. The other provinces did not see large increases in MAPE, though the 1-month prediction scores were already poor. MAPE rose from 29.3% to 42.4% in Lào Cai, and from 407% to 418% in Điện Biên, for one to three month ahead forecasts. The LSTM-ATT model failed to identify the one outbreak in Lào Cai, and there were no other outbreaks to evaluate on. The model's ability to correctly identify normal months, however, remained strong for long-term detections. These findings support the use of LSTM-ATT as a high performance long-term predictor of diarrhoea incidence in Thái Bình and Cao Bằng, though performance is suboptimal in other provinces.

There have been very few models developed for long-term forecasting of diarrhoea incidence that have been published in the literature for comparison. To the best of our knowledge, there have been no such climate-based models developed, nor any to produce MAPE values as low as 10.8% for 3-month ahead predictions.[1] Medina et al. (2007) employed a univariate multiplicative Holt-Winters model to forecast diarrhoea incidence in Niono, Mali. In contrast to our study, diarrhoea incidence was split into four age groups of 0–11 months, 1–4 years, 5–

[1] Based on PubMed and IEEE Xplore searches for ((diarrhoea[Title/Abstract]) OR (diarrhea[Title/Abstract])) AND ((forecasting[Title/Abstract]) OR (prediction[Title/Abstract])), and other works come across in literature review.

15 years, and >15 years old. Across the age groups, MAPE values ranged from 23.2% to 42.4% for predictions 2-months ahead, and from 22.8 to 43.5% for predictions 3-months ahead. Our findings show much lower error scores in some provinces. Furthermore, the long-term deep learning models developed here appear to be novel developments for long-term diarrhoea forecasting in Vietnam. Deployment of these models, therefore, has the potential for significant impact in mitigating morbidity associated with diarrhoea in Vietnam and elsewhere.

4.3 Study Limitations

The main limitations of this study were related to the predictors used, the reliability of data, and the impact of dengue virus serotypes on DF prediction. Firstly, we focused on climate factors (and influenza rates) as prediction variables, however there are many alternative factors affecting disease transmission For example, some of these factors include mosquito density, human travel metrics, population immunity, temporal data on water quality, and public health programs addressing DF and diarrhoea prevention and control. However, most of these suffer from limited data availability in Vietnam. Secondly, there were missing data in both the climate and case datasets, which may have negatively impacted model performance. Zero counts were assumed to be real values, but these could have represented missing data in addition to NAs. Finally, we did not attempt to model immunity or dengue virus serotype distribution which could impact model performance. Circulation of a new serotype in a region is believed to cause outbreaks of symptomatic DF cases every few years, and previous works have noted difficulty in modelling these spikes (Bett et al., 2019). During the three year test period, the DF models

in this study have shown reasonable accuracy in forecasting these multi-annual spikes. Still, data on serotype distribution may improve forecasts further.

5. Conclusion

A collection of machine learning models was implemented for forecasting DF and diarrhoeal disease incidence in Vietnam, with further analyses conducted on the highest performing models. DF prediction was evaluated across 20 provinces throughout Vietnam, covering a wide spread of different geographical and climate conditions. This facilitated a stringent assessment of our models, and it provided a broader evaluation compared to previous studies using fewer climate variables and fewer provinces (Pham et al., 2020; Phung et al., 2015b). The addition of an attention mechanism to the LSTM model decreased errors in most provinces for DF and diarrhoea forecasting, and the LSTM-ATT model outperformed competing forecast techniques overall. LSTM-ATT had the lowest MAE-based ranking for diarrhoea forecasting too, though there were only five evaluation provinces, and CNN had the lowest RMSE-based ranking. The models developed here provide the groundwork for early-warning systems for DF and diarrhoea incidence in Vietnam, and may contribute to reductions in national morbidity and mortality. Given the promising results obtained here, further studies developing LSTM-ATT and CNN models could be impactful in reducing the global burden of other climate-sensitive infectious diseases. Additional research should be carried out on developing and improving provincial level models, given the large variation in model performance across locations. This variation also suggests different machine learning techniques may be better for different

provinces. Thus, the development of superensemble methods for DF and diarrhoea forecasting in Vietnam may be able to provide models with more consistent accuracy across provinces.

Data and Code Availability

The population data used to calculate disease incidence rates per province per year are publicly available from the General Statistics Office of Vietnam at https://www.gso.gov.vn/en/population/. The climate and disease data were obtained for a fee from IMHEN and NIHE, respectively. Restrictions apply to the availability of the data, which is available from the author with the permission of the respective institutions. Alternatively, data can be requested directly from IMHEN and NIHE.

The code used for this project is provided in a public GitHub repository, containing the code itself and a brief overview of the function of each file. This is available at the following link: https://github.com/mullach/climate-sensitive-diseases

Supplementary Material

Table S1: Numbers of layers and hidden sizes for LSTM, LSTM-ATT, and Transformer for all provinces. Listed in the format layers – hidden sizes. LSTM = Long Short-Term Memory, LSTM-ATT = Attention mechanism-enhanced LSTM.

Province	LSTM	LSTM-ATT	Transformer
Hà Nội	3 - 128	3 - 512	2 – 512
Hải Phòng	3 - 512	3 - 256	3 – 128
Quảng Nam	2 - 384	2 - 512	4 – 256
Quảng Ngãi	3 - 128	4 - 256	3 – 512
Kon Tum	3 - 384	2 - 256	3 – 256
Phú Yên	4 - 512	2 - 128	3 - 384
Ninh Thuận	4 - 512	2 - 384	2 – 256
Bình Thuận	3 - 256	3 - 256	4 - 128
Tây Ninh	4 - 256	3 - 384	3 – 128
Bình Phước	2 - 512	2 - 256	4 – 256
An Giang	4 - 384	3 - 384	4 – 128
Tiền Giang	2 - 512	3 - 128	3 – 384
Cần Thơ	4 - 512	3 - 512	2 – 128
Trà Vinh	4 - 512	4 – 256	2 – 384
Kiên Giang	3 - 512	3 – 256	2 – 256
Sóc Trăng	2 - 512	2 - 256	2 – 384
Bạc Liêu	4 - 128	3 - 128	2 – 384
Cà Mau	3 - 384	3 - 256	2 - 384
Gia Lai	3 - 256	3 - 256	2 - 384
Nam Định	2 - 512	2 - 256	2 – 512
Thái Bình	3 - 256	2 - 256	3 – 256
Quảng Ninh	4 - 256	4 - 384	3 - 128

Table S2: Model hyperparameters for diarrhoea predictions. LSTM = Long Short-Term Memory, LSTM-ATT = Attention mechanism-enhanced LSTM. CNN = Convolutional Neural Network.

Model	Province	n Features	Batch Size	Lookback Window	Epochs	Learning Rate	Hidden Size	n Layers	Num. Filters	Dropout Rate
LSTM	Điện Biên'	3	16	3	120	0.00303	23	1	-	-
	Thái Bình	3	16	3	40	0.0045	140	4	-	-
	Lào Cai	3	16	3	70	0.000133	70	2	-	-
	Kon Tum	3	16	4	140	0.000769	11	9	-	-
	Cao Bằng	3	16	3	40	0.00515	8	1	-	-
LSTM-ATT	Điện Biên'	3	16	3	360	0.000905	41	2	-	-
	Thái Bình	3	16	3	360	0.000273	79	3	-	-
	Lào Cai	3	16	3	480	0.000243	25	4	-	-
	Kon Tum	3	16	2	180	0.00879	30	4	-	-
	Cao Bằng	3	16	2	410	0.00675	15	6	-	-
CNN	Điện Biên'	3	16	3	70	0.00813	-	-	[16, 32, 64]	0.685
	Thái Bình	3	16	3	440	0.00646	-	-	[100, 100, 100]	0.477
	Lào Cai	3	16	3	130	0.00609	-	-	[16, 32, 64]	0.7
	Kon Tum	3	16	3	140	0.000129	-	-	[32, 64, 128]	0.732
	Cao Bằng	3	16	3	220	0.00696	-	-	[64, 64, 64]	0.49

Table S3: Shapiro-Wilk normality test results.

Variable	Shapiro statistic	P-value
Influenza cases	0.716	0.00E+00
Dengue fever cases	0.385	0.00E+00
Diarrhoea cases	0.628	0.00E+00
Total evaporation	0.946	0.00E+00
Total rainfall	0.779	0.00E+00
Max daily rainfall	0.754	0.00E+00
Number of raining days	0.939	0.00E+00
Average temperature	0.907	0.00E+00
Max average temperature	0.915	0.00E+00
Min average temperature	0.905	0.00E+00
Max absolute temperature	0.970	1.40E-45
Min absolute temperature	0.893	0.00E+00
Average humidity	0.982	6.19E-38
Min humidity	0.991	2.18E-28
Number of sunshine hours	0.992	8.92E-27
Influenza rates	0.664	0.00E+00
Dengue fever rates	0.360	0.00E+00
Diarrhoea rates	0.706	0.00E+00

Table S4: Kruskal-Wallis H-test results for differences in climate and disease variables between regions in Vietnam. Significant values indicate a difference in population medians between north, central, and south Vietnam (*: $p < 0.05$, **: $p < 0.01$).

Variable	H Statistic	P-value
Dengue fever rates	6065.968042	0.00E+00***
Diarrhoea rates	183.0981062	1.74E-40***
Total rainfall	0.716356439	6.99E-01
Average temperature	1960.319729	0.00E+00***
Average humidity	149.5476477	3.36E-33***
Total evaporation	1097.991641	3.75E-239***
Total sunshine hours	3253.918943	0.00E+00***

Table S5: Pairwise differences in climate and disease variables between regions in Vietnam. P-values generated from Dunn's post-hoc test. Significance thresholds have been adjusted using Bonferroni correction for 54 comparisons (*: $p < 0.05$, **: $p < 0.01$).

Variable		P-value		
	Region	**Central**	**North**	**South**
Dengue fever rates	Central	1.00E+00	0.00E+00***	5.04E-145***
	North	0.00E+00***	1.00E+00	0.00E+00***
	South	5.04E-145***	0.00E+00***	1.00E+00
Diarrhoea rates	Central	1.00E+00	8.04E-37***	2.13E-02
	North	8.04E-37***	1.00E+00	4.07E-24***
	South	2.13E-02	4.07E-24***	1.00E+00
Average temperature	Central	1.00E+00	1.50E-146***	8.22E-47***
	North	1.50E-146***	1.00E+00	0.00E+00***
	South	8.22E-47***	0.00E+00***	1.00E+00
Average humidity	Central	1.00E+00	9.51E-14***	8.56E-04*
	North	9.51E-14***	1.00E+00	1.97E-29***
	South	8.56E-04*	1.97E-29***	1.00E+00
Total evaporation	Central	1.00E+00	8.18E-109***	2.36E-12***
	North	8.18E-109***	1.00E+00	1.30E-200***
	South	2.36E-12***	1.30E-200***	1.00E+00
Total sunshine hours	Central	1.00E+00	1.73E-209***	1.46E-106***
	North	1.73E-209***	1.00E+00	0.00E+00***
	South	1.46E-106***	0.00E+00***	1.00E+00

Table S6: Mean absolute errors for all prediction models in 20 Vietnamese provinces. Values are colour-coded for each province separately from the lowest value (darker green) to the median value (yellow) to the highest value (darker red). LSTM = long short-term memory. LSTM-ATT = attention mechanism-enhanced LSTM. CNN = convolutional neural network. Poisson = Poisson regression. XGBoost = Extreme Gradient Boosting. SVR = Support Vector Regressor with Radial Basis Kernel. SVR-L = Support Vector Regressor with Linear Kernel. SARIMA = Seasonal Autoregressive Integrated Moving Average.

Province	Mean Absolute Error for Each Model								
	LSTM	LSTM-ATT	CNN	TF	Poisson	XGB	SVR	SVR-L	SARIMA
Hà Nội	4.926	3.457	5.065	5.695	8.397	7.199	9.077	9.542	8.637
Hải Phòng	0.276	0.366	0.538	0.702	0.817	0.434	5.196	7.838	2.541
Quảng Ninh	0.652	0.614	1.223	0.560	1.325	0.876	2.945	3.973	0.786
Nam Định	0.556	0.492	0.654	0.748	0.796	0.806	1.229	1.428	0.728
Thái Bình	0.412	0.432	0.428	0.468	0.498	0.420	0.664	0.803	0.522
Quảng Nam	3.766	4.116	4.039	8.353	8.730	8.216	9.567	11.802	10.505
Quảng Ngãi	6.699	6.579	6.183	5.913	9.442	6.739	24.494	36.921	7.112
Phú Yên	6.604	7.342	6.433	10.167	13.429	11.923	15.608	17.670	18.062
Ninh Thuận	3.733	2.813	3.875	5.351	15.816	17.633	17.566	9.028	5.589
Bình Thuận	6.606	6.495	6.300	9.692	9.929	7.755	11.225	11.898	7.280
Tây Ninh	4.405	2.837	5.218	5.305	5.517	6.622	5.460	8.220	5.585
Bình Phước	5.020	5.353	6.846	7.546	10.957	10.042	14.780	13.715	16.440
An Giang	4.462	3.006	2.769	3.747	8.476	7.057	6.762	7.021	9.423
Tiền Giang	3.845	3.371	6.589	4.876	15.919	13.528	10.893	14.204	10.671
Cần Thơ	2.611	1.884	2.911	4.469	6.725	4.864	16.782	27.370	8.148
Trà Vinh	3.143	2.702	3.528	4.005	9.376	9.435	11.766	11.692	7.984
Kiên Giang	1.848	2.093	3.537	3.110	13.859	12.334	14.397	14.652	3.765
Sóc Trăng	4.393	4.540	3.084	3.304	10.683	10.326	10.310	10.283	36.243
Bạc Liêu	2.870	2.160	2.008	2.207	11.494	9.399	9.142	8.897	19.599
Cà Mau	3.553	2.935	4.582	5.710	12.015	11.213	13.103	14.381	16.263

Table S7: Selected features for all provinces for dengue fever models.

Province	Feature 1	Feature 2
Hà Nội	Max average temperature	Min absolute temperature
Hải Phòng	Max daily rainfall	Max average temperature
Quảng Nam	Total rainfall	Average temperature
Quảng Ngãi	Number of raining days	Average humidity
Kon Tum	Total rainfall	Min average temperature
Phú Yên	Total evaporation	Number of raining days
Ninh Thuận	Total evaporation	Number of raining days
Bình Thuận	Total rainfall	Max average temperature
Tây Ninh	Max daily rainfall	Average temperature
Bình Phước	Total evaporation	Max absolute temperature
An Giang	Max daily rainfall	Number of raining days
Tiền Giang	Max daily rainfall	Average humidity
Cần Thơ	Average temperature	Number of sunshine hours
Trà Vinh	Max daily rainfall	Max absolute temperature
Kiên Giang	Max average temperature	Max absolute temperature
Sóc Trăng	Number of raining days	Number of sunshine hours
Bạc Liêu	Max daily rainfall	Max absolute temperature
Cà Mau	Total rainfall	Max daily rainfall
Gia Lai	Total evaporation	Average humidity
Nam Định	Min absolute temperature	Number of sunshine hours
Thái Bình	Max average temperature	Number of sunshine hours
Quảng Ninh	Number of raining days	Number of sunshine hours

Table S8: Mean absolute percentage errors for LSTM, LSTM-ATT, and CNN in 20 Vietnamese provinces. Values are colour-coded for each province separately from the lowest value (darker green) to the median value (yellow) to the highest value (darker red). LSTM = long short-term memory. LSTM-ATT = attention mechanism-enhanced LSTM. CNN = convolutional neural network.

Province	MAPE for Each Model		
	LSTM	LSTM-ATT	CNN
An Giang	0.741	0.384	0.405
Cần Thơ	0.770	0.483	1.280
Tiền Giang	0.645	0.510	1.323
Sóc Trăng	0.684	0.532	0.517
Tây Ninh	1.281	0.583	2.104
Cà Mau	1.210	0.664	1.314
Bạc Liêu	1.602	0.780	1.248
Bình Thuận	1.229	0.896	1.568
Bình Phước	0.879	0.981	0.978
Phú Yên	0.734	1.053	0.744
Hà Nội	7.881	1.728	6.305
Trà Vinh	3.337	1.993	4.460
Ninh Thuận	2.186	2.725	3.089
Quảng Ngãi	5.513	4.144	4.370
Nam Định	6.29E+15	3.05E+15	7.96E+15
Quảng Nam	2.81E+15	3.48E+15	4.20E+15
Kon Tum	1.47E+16	4.25E+15	3.90E+16
Thái Bình	4.95E+15	4.72E+15	4.00E+15
Quảng Ninh	9.49E+15	7.79E+15	8.64E+15
Hải Phòng	9.62E+15	9.14E+15	1.60E+16

Table S9: Selected features for all provinces for diarrhoea prediction models.

Province	Feature 1	Feature 2
Điện Biên	Max daily rainfall	Influenza rates
Thái Bình	Total rainfall	Number of sunshine hours
Lào Cai	Min average temperature	Number of sunshine hours
Kon Tum	Min absolute temperature	Influenza rates
Cao Bằng	Min average temperature	Influenza rates

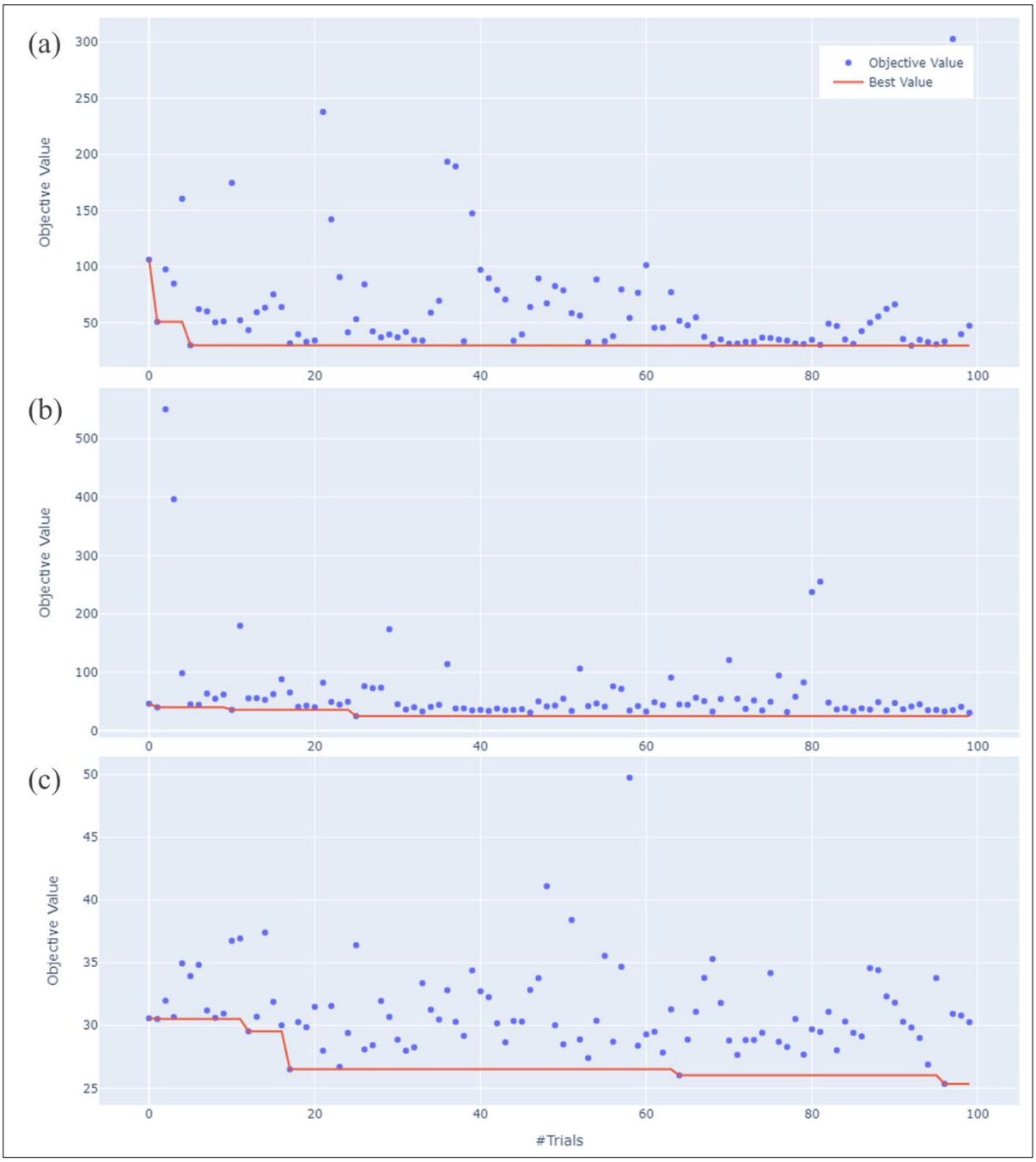

Figure S1: Optimisation histories for deep learning models. Root mean square errors are shown for each trial as blue dots, and the red line represents the minimum value as the trials progress. Histories are displayed for (a) Long Short-Term Memory (LSTM), (b) Attention mechanism-enhanced LSTM, and (c) Convolutional Neural Network models. N.B. objective value y-scales vary between plots.

References

Abdullahi, T., Nitschke, G., 2021. Predicting Disease Outbreaks with Climate Data, in: 2021 IEEE Congress on Evolutionary Computation (CEC). Presented at the 2021 IEEE Congress on Evolutionary Computation (CEC), pp. 989–996. https://doi.org/10.1109/CEC45853.2021.9504740

Aguas, R., Dorigatti, I., Coudeville, L., Luxemburger, C., Ferguson, N.M., 2019. Cross-serotype interactions and disease outcome prediction of dengue infections in Vietnam. Sci. Rep. 9, 9395. https://doi.org/10.1038/s41598-019-45816-6

Akiba, T., Sano, S., Yanase, T., Ohta, T., Koyama, M., 2019. Optuna: A Next-generation Hyperparameter Optimization Framework, in: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD '19. Association for Computing Machinery, New York, NY, USA, pp. 2623–2631. https://doi.org/10.1145/3292500.3330701

Albawi, S., Mohammed, T.A., Al-Zawi, S., 2017. Understanding of a convolutional neural network, in: 2017 International Conference on Engineering and Technology (ICET). Presented at the 2017 International Conference on Engineering and Technology (ICET), pp. 1–6. https://doi.org/10.1109/ICEngTechnol.2017.8308186

Ali, M., Kim, D.R., Yunus, M., Emch, M., 2013. Time Series Analysis of Cholera in Matlab, Bangladesh, during 1988-2001. J. Health Popul. Nutr. 31, 11–19.

Anders, K.L., Thompson, C.N., Thuy, N.T.V., Nguyet, N.M., Tu, L.T.P., Dung, T.T.N., Phat, V.V., Van, N.T.H., Hieu, N.T., Tham, N.T.H., Ha, P.T.T., Lien, L.B., Chau, N.V.V., Baker, S., Simmons, C.P., 2015. The epidemiology and aetiology of diarrhoeal disease in infancy in southern Vietnam: a birth cohort study. Int. J. Infect. Dis. 35, 3–10. https://doi.org/10.1016/j.ijid.2015.03.013

Awad, M., Khanna, R., 2015. Support Vector Regression, in: Awad, M., Khanna, R. (Eds.), Efficient Learning Machines: Theories, Concepts, and Applications for Engineers and System Designers. Apress, Berkeley, CA, pp. 67–80. https://doi.org/10.1007/978-1-4302-5990-9_4

Bahdanau, D., Cho, K., Bengio, Y., 2016. Neural Machine Translation by Jointly Learning to Align and Translate. ArXiv14090473 Cs Stat.

Barrera, R., Amador, M., MacKay, A.J., 2011. Population Dynamics of Aedes aegypti and Dengue as Influenced by Weather and Human Behavior in San Juan, Puerto Rico. PLoS Negl. Trop. Dis. 5, e1378. https://doi.org/10.1371/journal.pntd.0001378

Bengio, Y., Courville, A., Vincent, P., 2013. Representation Learning: A Review and New Perspectives. IEEE Trans. Pattern Anal. Mach. Intell. 35, 1798–1828. https://doi.org/10.1109/TPAMI.2013.50

Bergstra, J., Bengio, Y., 2012. Random Search for Hyper-Parameter Optimization. J. Mach. Learn. Res. 13, 281–305.

Bett, B., Grace, D., Lee, H.S., Lindahl, J., Nguyen-Viet, H., Phuc, P.-D., Quyen, N.H., Tu, T.A., Phu, T.D., Tan, D.Q., Nam, V.S., 2019. Spatiotemporal analysis of historical records (2001–2012) on dengue fever in Vietnam and development of a statistical model for forecasting risk. PLOS ONE 14, e0224353. https://doi.org/10.1371/journal.pone.0224353

Bharaj, P., Chahar, H.S., Pandey, A., Diddi, K., Dar, L., Guleria, R., Kabra, S.K., Broor, S., 2008. Concurrent infections by all four dengue virus serotypes during an outbreak of dengue in 2006 in Delhi, India. Virol. J. 5, 1. https://doi.org/10.1186/1743-422X-5-1

Bhatt, S., Gething, P.W., Brady, O.J., Messina, J.P., Farlow, A.W., Moyes, C.L., Drake, J.M., Brownstein, J.S., Hoen, A.G., Sankoh, O., Myers, M.F., George, D.B., Jaenisch, T.,

Wint, G.R.W., Simmons, C.P., Scott, T.W., Farrar, J.J., Hay, S.I., 2013. The global distribution and burden of dengue. Nature 496, 504–507. https://doi.org/10.1038/nature12060

Bich, T.H., Quang, L.N., Thanh Ha, L.T., Duc Hanh, T.T., Guha-Sapir, D., 2011. Impacts of flood on health: epidemiologic evidence from Hanoi, Vietnam. Glob. Health Action 4, 6356. https://doi.org/10.3402/gha.v4i0.6356

Black, R.E., Brown, K.H., Becker, S., Alim, A.R.M.A., Merson, M.H., 1982. Contamination of weaning foods and transmission of enterotoxigenic Escherichia coli diarrhoea in children in rural Bangladesh. Trans. R. Soc. Trop. Med. Hyg. 76, 259–264. https://doi.org/10.1016/0035-9203(82)90292-9

Blaustein, R.A., Pachepsky, Y., Hill, R.L., Shelton, D.R., Whelan, G., 2013. Escherichia coli survival in waters: Temperature dependence. Water Res. 47, 569–578. https://doi.org/10.1016/j.watres.2012.10.027

Boateng, I., 2012. GIS assessment of coastal vulnerability to climate change and coastal adaption planning in Vietnam. J. Coast. Conserv. 16, 25–36. https://doi.org/10.1007/s11852-011-0165-0

Brady, O.J., Smith, D.L., Scott, T.W., Hay, S.I., 2015. Dengue disease outbreak definitions are implicitly variable. Epidemics 11, 92–102. https://doi.org/10.1016/j.epidem.2015.03.002

Cai, M., Pipattanasomporn, M., Rahman, S., 2019. Day-ahead building-level load forecasts using deep learning vs. traditional time-series techniques. Appl. Energy 236, 1078–1088. https://doi.org/10.1016/j.apenergy.2018.12.042

Carlton, E.J., Woster, A.P., DeWitt, P., Goldstein, R.S., Levy, K., 2016. A systematic review and meta-analysis of ambient temperature and diarrhoeal diseases. Int. J. Epidemiol. 45, 117–130. https://doi.org/10.1093/ije/dyv296

CARTO, 2019. List of available basemaps in CARTO [WWW Document]. CARTO. URL https://carto.com/help/building-maps/basemap-list/ (accessed 8.17.21).

Checkley, W., Epstein, L.D., Gilman, R.H., Figueroa, D., Cama, R.I., Patz, J.A., Black, R.E., 2000. Effects of EI Niño and ambient temperature on hospital admissions for diarrhoeal diseases in Peruvian children. The Lancet 355, 442–450. https://doi.org/10.1016/S0140-6736(00)82010-3

Chen, T., Guestrin, C., 2016. XGBoost: A Scalable Tree Boosting System. Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min. 785–794. https://doi.org/10.1145/2939672.2939785

Chen, Y., Ong, J.H.Y., Rajarethinam, J., Yap, G., Ng, L.C., Cook, A.R., 2018. Neighbourhood level real-time forecasting of dengue cases in tropical urban Singapore. BMC Med. 16, 129. https://doi.org/10.1186/s12916-018-1108-5

Cheng, J., Bambrick, H., Yakob, L., Devine, G., Frentiu, F.D., Toan, D.T.T., Thai, P.Q., Xu, Z., Hu, W., 2020. Heatwaves and dengue outbreaks in Hanoi, Vietnam: New evidence on early warning. PLoS Negl. Trop. Dis. 14, e0007997. https://doi.org/10.1371/journal.pntd.0007997

Colón-González, F.J., Bastos, L.S., Hofmann, B., Hopkin, A., Harpham, Q., Crocker, T., Amato, R., Ferrario, I., Moschini, F., James, S., Malde, S., Ainscoe, E., Nam, V.S., Tan, D.Q., Khoa, N.D., Harrison, M., Tsarouchi, G., Lumbroso, D., Brady, O.J., Lowe, R., 2021. Probabilistic seasonal dengue forecasting in Vietnam: A modelling study using superensembles. PLOS Med. 18, e1003542. https://doi.org/10.1371/journal.pmed.1003542

Colón-González, F.J., Lake, I.R., Bentham, G., 2011. Climate Variability and Dengue Fever in Warm and Humid Mexico. Am. J. Trop. Med. Hyg. 84, 757–763. https://doi.org/10.4269/ajtmh.2011.10-0609

Ding, G., Li, Xiaomei, Li, Xuewen, Zhang, B., Jiang, B., Li, D., Xing, W., Liu, Q., Liu, X., Hou, H., 2019. A time-trend ecological study for identifying flood-sensitive infectious diseases in Guangxi, China from 2005 to 2012. Environ. Res. 176, 108577. https://doi.org/10.1016/j.envres.2019.108577

Ding, G., Zhang, Y., Gao, L., Ma, W., Li, X., Liu, J., Liu, Q., Jiang, B., 2013. Quantitative Analysis of Burden of Infectious Diarrhea Associated with Floods in Northwest of Anhui Province, China: A Mixed Method Evaluation. PLOS ONE 8, e65112. https://doi.org/10.1371/journal.pone.0065112

Do, T.T.T., Martens, P., Luu, N.H., Wright, P., Choisy, M., 2014. Climatic-driven seasonality of emerging dengue fever in Hanoi, Vietnam. BMC Public Health 14, 1078. https://doi.org/10.1186/1471-2458-14-1078

D'souza, R.M., Hall, G., Becker, N.G., 2008. Climatic factors associated with hospitalizations for rotavirus diarrhoea in children under 5 years of age. Epidemiol. Infect. 136, 56–64. https://doi.org/10.1017/S0950268807008229

Eisen, L., Monaghan, A.J., Lozano-Fuentes, S., Steinhoff, D.F., Hayden, M.H., Bieringer, P.E., 2014. The Impact of Temperature on the Bionomics of Aedes (Stegomyia) aegypti, With Special Reference to the Cool Geographic Range Margins. J. Med. Entomol. 51, 496–516. https://doi.org/10.1603/ME13214

El-Senousy, W.M., Osman, G.A., Melegy, A., 2014. Survival of adenovirus, rotavirus, Hepatitis A virus, pathogenic bacteria and bacterial indicators in ground water. World Appl. Sci. J. 29, 337–348. https://doi.org/10.5829/idosi.wasj.2014.29.03.13849

European Medicines Agency, 2020. Dengvaxia [WWW Document]. Eur. Med. Agency. URL https://www.ema.europa.eu/en/medicines/human/EPAR/dengvaxia (accessed 9.11.21).

Fang, X., Ai, J., Liu, W., Ji, H., Zhang, X., Peng, Z., Wu, Y., Shi, Y., Shen, W., Bao, C., 2019. Epidemiology of infectious diarrhoea and the relationship with etiological and meteorological factors in Jiangsu Province, China. Sci. Rep. 9, 19571. https://doi.org/10.1038/s41598-019-56207-2

Fang, X., Liu, W., Ai, J., He, M., Wu, Y., Shi, Y., Shen, W., Bao, C., 2020. Forecasting incidence of infectious diarrhea using random forest in Jiangsu Province, China. BMC Infect. Dis. 20, 222. https://doi.org/10.1186/s12879-020-4930-2

FAO, 2011. AQUASTAT Country profile — Viet Nam. Food and Agriculture Organization of the United Nations (FAO), Rome, Italy.

General Statistics Office of Vietnam, 2021a. Average population by province, sex and residence [WWW Document]. Gen. Stat. Off. Vietnam. URL https://www.gso.gov.vn/en/px-web/ (accessed 8.14.21).

General Statistics Office of Vietnam, 2021b. Health, Culture, Sport, Living standards, Social order, Safety and Environment [WWW Document]. Gen. Stat. Off. Vietnam. URL https://www.gso.gov.vn/en/health-culture-sport-living-standards-social-order-safety-and-environment/ (accessed 9.7.21).

Gilca, R., Fortin, É., Frenette, C., Longtin, Y., Gourdeau, M., 2012. Seasonal Variations in Clostridium difficile Infections Are Associated with Influenza and Respiratory Syncytial Virus Activity Independently of Antibiotic Prescriptions: a Time Series Analysis in Québec, Canada. Antimicrob. Agents Chemother. 56, 639–646. https://doi.org/10.1128/AAC.05411-11

Graves, A., Schmidhuber, J., 2005. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. Neural Netw., IJCNN 2005 18, 602–610. https://doi.org/10.1016/j.neunet.2005.06.042

Halstead, S.B., 2014. Dengue Antibody-Dependent Enhancement: Knowns and Unknowns. Microbiol. Spectr. 2. https://doi.org/10.1128/microbiolspec.AID-0022-2014

Higa, Y., Yen, N.T., Kawada, H., Son, T.H., Hoa, N.T., Takagi, M., 2010. Geographic distribution of Aedes aegypti and Aedes albopictus collected from used tires in Vietnam. J. Am. Mosq. Control Assoc. 26, 1–9. https://doi.org/10.2987/09-5945.1

Hii, Y.L., Zhu, H., Ng, N., Ng, L.C., Rocklöv, J., 2012. Forecast of Dengue Incidence Using Temperature and Rainfall. PLoS Negl. Trop. Dis. 6, e1908. https://doi.org/10.1371/journal.pntd.0001908

Hunter, J.D., 2007. Matplotlib: A 2D Graphics Environment. Comput. Sci. Eng. 9, 90–95. https://doi.org/10.1109/MCSE.2007.55

Huyen, D.T.T., Hong, D.T., Trung, N.T., Hoa, T.T.N., Oanh, N.K., Thang, H.V., Thao, N.T.T., Hung, D.M., Iijima, M., Fox, K., Grabovac, V., Heffelfinger, J., Batmunkh, N., Anh, D.D., 2018. Epidemiology of acute diarrhea caused by rotavirus in sentinel surveillance sites of Vietnam, 2012–2015. Vaccine, Rotavirus Surveillance, Safety and Economic Data before Vaccine Introduction: a Global Perspective from the World Health Organization Global Rotavirus Surveillance Network 36, 7894–7900. https://doi.org/10.1016/j.vaccine.2018.05.008

Hyndman, R., 2010. The ARIMAX model muddle [WWW Document]. Rob J Hyndman. URL https://robjhyndman.com/hyndsight/arimax/ (accessed 9.5.21).

Institute of Strategy and Policy on Natural Resources and Environment, 2009. Viet Nam Assessment Report on Climate Change (VARCC). Institute of Strategy and Policy on Natural Resources and Environmen (ISPONRE), Viet Nam.

International Organization for Standardization, 2020. VN - Viet Nam [WWW Document]. ISO Online Brows. Platf. URL https://www.iso.org/obp/ui/#iso:code:3166:VN (accessed 9.5.21).

Isenbarger, D.W., Hien, B.T., Ha, H.T., Ha, T.T., Bodhidatta, L., Pang, L.W., Cam, P.D., 2001. Prospective study of the incidence of diarrhoea and prevalence of bacterial pathogens in a cohort of Vietnamese children along the Red River. Epidemiol. Infect. 127, 229–236. https://doi.org/10.1017/S0950268801005933

Islam, M.S., Sharker, M.A.Y., Rheman, S., Hossain, S., Mahmud, Z.H., Islam, M.S., Uddin, A.M.K., Yunus, M., Osman, M.S., Ernst, R., Rector, I., Larson, C.P., Luby, S.P., Endtz, H.P., Cravioto, A., 2009. Effects of local climate variability on transmission dynamics of cholera in Matlab, Bangladesh. Trans. R. Soc. Trop. Med. Hyg. 103, 1165–1170. https://doi.org/10.1016/j.trstmh.2009.04.016

Jia, D., Wang, R., Xu, C., Yu, Z., 2016. QIM: Quantifying Hyperparameter Importance for Deep Learning, in: Gao, G.R., Qian, D., Gao, X., Chapman, B., Chen, W. (Eds.), Network and Parallel Computing, Lecture Notes in Computer Science. Springer International Publishing, Cham, pp. 180–188. https://doi.org/10.1007/978-3-319-47099-3_15

Jia, W., Wan, Y., Li, Y., Tan, K., Lei, W., Hu, Y., Ma, Z., Li, X., Xie, G., 2019. Integrating Multiple Data Sources and Learning Models to Predict Infectious Diseases in China. AMIA Summits Transl. Sci. Proc. 2019, 680–685.

Jordahl, K., Bossche, J.V.D., Fleischmann, M., Wasserman, J., McBride, J., Gerard, J., Tratner, J., Perry, M., Badaracco, A.G., Farmer, C., Hjelle, G.A., Snow, A.D., Cochran, M., Gillies, S., Culbertson, L., Bartos, M., Eubank, N., Maxalbert, Bilogur, A., Rey, S., Ren, C., Arribas-Bel, D., Wasser, L., Wolf, L.J., Journois, M., Wilson, J., Greenhall, A., Holdgraf, C., Filipe, Leblanc, F., 2020. geopandas/geopandas: v0.9.0. Zenodo. https://doi.org/10.5281/ZENODO.3946761

Kam, H.J., Choi, S., Cho, J.P., Min, Y.G., Park, R.W., 2010. Acute diarrheal syndromic surveillance: effects of weather and holidays. Appl. Clin. Inform. 1, 79–95. https://doi.org/10.4338/ACI-2009-12-RA-0024

Kırbaş, İ., Sözen, A., Tuncer, A.D., Kazancıoğlu, F.Ş., 2020. Comparative analysis and forecasting of COVID-19 cases in various European countries with ARIMA, NARNN and LSTM approaches. Chaos Solitons Fractals 138, 110015. https://doi.org/10.1016/j.chaos.2020.110015

Kotloff, K.L., Nataro, J.P., Blackwelder, W.C., Nasrin, D., Farag, T.H., Panchalingam, S., Wu, Y., Sow, S.O., Sur, D., Breiman, R.F., Faruque, A.S., Zaidi, A.K., Saha, D., Alonso, P.L., Tamboura, B., Sanogo, D., Onwuchekwa, U., Manna, B., Ramamurthy, T., Kanungo, S., Ochieng, J.B., Omore, R., Oundo, J.O., Hossain, A., Das, S.K., Ahmed, S., Qureshi, S., Quadri, F., Adegbola, R.A., Antonio, M., Hossain, M.J., Akinsola, A., Mandomando, I., Nhampossa, T., Acácio, S., Biswas, K., O'Reilly, C.E., Mintz, E.D., Berkeley, L.Y., Muhsen, K., Sommerfelt, H., Robins-Browne, R.M., Levine, M.M., 2013. Burden and aetiology of diarrhoeal disease in infants and young children in developing countries (the Global Enteric Multicenter Study, GEMS): a prospective, case-control study. The Lancet 382, 209–222. https://doi.org/10.1016/S0140-6736(13)60844-2

Lee, H.S., Ha Hoang, T.T., Pham-Duc, P., Lee, M., Grace, D., Phung, D.C., Thuc, V.M., Nguyen-Viet, H., 2017a. Seasonal and geographical distribution of bacillary dysentery (shigellosis) and associated climate risk factors in Kon Tum Province in Vietnam from 1999 to 2013. Infect. Dis. Poverty 6, 113. https://doi.org/10.1186/s40249-017-0325-z

Lee, H.S., Nguyen-Viet, H., Nam, V.S., Lee, M., Won, S., Duc, P.P., Grace, D., 2017b. Seasonal patterns of dengue fever and associated climate factors in 4 provinces in Vietnam from 1994 to 2013. BMC Infect. Dis. 17, 218. https://doi.org/10.1186/s12879-017-2326-8

Lega, J., Brown, H.E., Barrera, R., 2017. Aedes aegypti (Diptera: Culicidae) Abundance Model Improved With Relative Humidity and Precipitation-Driven Egg Hatching. J. Med. Entomol. 54, 1375–1384. https://doi.org/10.1093/jme/tjx077

Liu, K., Zhang, M., Xi, G., Deng, A., Song, T., Li, Q., Kang, M., Yin, L., 2020. Enhancing fine-grained intra-urban dengue forecasting by integrating spatial interactions of human movements between urban regions. PLoS Negl. Trop. Dis. 14, e0008924. https://doi.org/10.1371/journal.pntd.0008924

Liu-Helmersson, J., Stenlund, H., Wilder-Smith, A., Rocklöv, J., 2014. Vectorial Capacity of Aedes aegypti: Effects of Temperature and Implications for Global Dengue Epidemic Potential. PLoS ONE 9. https://doi.org/10.1371/journal.pone.0089783

Lowe, R., Gasparrini, A., Meerbeeck, C.J.V., Lippi, C.A., Mahon, R., Trotman, A.R., Rollock, L., Hinds, A.Q.J., Ryan, S.J., Stewart-Ibarra, A.M., 2018. Nonlinear and delayed impacts of climate on dengue risk in Barbados: A modelling study. PLOS Med. 15, e1002613. https://doi.org/10.1371/journal.pmed.1002613

Luong, M.-T., Pham, H., Manning, C.D., 2015. Effective Approaches to Attention-based Neural Machine Translation. ArXiv150804025 Cs.

Medina, D.C., Findley, S.E., Guindo, B., Doumbia, S., 2007. Forecasting Non-Stationary Diarrhea, Acute Respiratory Infection, and Malaria Time-Series in Niono, Mali. PLoS ONE 2, e1181. https://doi.org/10.1371/journal.pone.0001181

Mustafa, M.S., Rasotgi, V., Jain, S., Gupta, V., 2015. Discovery of fifth serotype of dengue virus (DENV-5): A new public health dilemma in dengue control. Med. J. Armed Forces India 71, 67–70. https://doi.org/10.1016/j.mjafi.2014.09.011

Nguyen, T.A., Vu, D.A., Van Vu, P., Nguyen, T.N., Pham, T.M., Nguyen, H.T.T., Trinh Le, H., Nguyen, T.V., Hoang, L.K., Vu, T.D., Nguyen, T.S., Luong, T.T., Trinh, N.P., Hens, L., 2017. Human ecological effects of tropical storms in the coastal area of Ky Anh (Ha Tinh, Vietnam). Environ. Dev. Sustain. 19, 745–767. https://doi.org/10.1007/s10668-016-9761-3

Nguyen, T.V., Le Van, P., Le Huy, C., Weintraub, A., 2004. Diarrhea Caused by Rotavirus in Children Less than 5 Years of Age in Hanoi, Vietnam. J. Clin. Microbiol. 42, 5745–5750. https://doi.org/10.1128/JCM.42.12.5745-5750.2004

Nguyen, T.V., Van, P.L., Huy, C.L., Gia, K.N., Weintraub, A., 2006. Etiology and epidemiology of diarrhea in children in Hanoi, Vietnam. Int. J. Infect. Dis. 10, 298–308. https://doi.org/10.1016/j.ijid.2005.05.009

Normile, D., 2013. Surprising New Dengue Virus Throws a Spanner in Disease Control Efforts. Science 342, 415–415. https://doi.org/10.1126/science.342.6157.415

Oh, E.J., Kim, J.M., Kim, J.K., 2021. Interrelationship between climatic factors and incidence of FBD caused by Clostridioides difficile toxin B, Clostridium perfringens, Campylobacter spp., and Escherichia coli O157:H7. Environ. Sci. Pollut. Res. 28, 44538–44546. https://doi.org/10.1007/s11356-021-13854-1

Okewu, E., Adewole, P., Sennaike, O., 2019. Experimental Comparison of Stochastic Optimizers in Deep Learning, in: Misra, S., Gervasi, O., Murgante, B., Stankova, E., Korkhov, V., Torre, C., Rocha, A.M.A.C., Taniar, D., Apduhan, B.O., Tarantino, E. (Eds.), Computational Science and Its Applications – ICCSA 2019, Lecture Notes in Computer Science. Springer International Publishing, Cham, pp. 704–715. https://doi.org/10.1007/978-3-030-24308-1_55

Onozuka, D., Hashizume, M., 2011. Weather variability and paediatric infectious gastroenteritis. Epidemiol. Infect. 139, 1369–1378. https://doi.org/10.1017/S0950268810002451

Parvez, S.M., Kwong, L., Rahman, M.J., Ercumen, A., Pickering, A.J., Ghosh, P.K., Rahman, M.Z., Das, K.K., Luby, S.P., Unicomb, L., 2017. Escherichia coli contamination of child complementary foods and association with domestic hygiene in rural Bangladesh. Trop. Med. Int. Health 22, 547–557. https://doi.org/10.1111/tmi.12849

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S., 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library, in: Advances in Neural Information Processing Systems 32. Curran Associates Inc., pp. 8024–8035.

PATH, 2019. Using innovation to combat diarrheal disease in Vietnam (Fact sheet). URL https://www.path.org/resources/using-innovation-combat-diarrheal-disease-vietnam/ (Accessed 18.8.21).

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, É., 2011. Scikit-learn: Machine Learning in Python. J. Mach. Learn. Res. 12, 2825–2830.

Perktold, Skipper, S., Taylor, J., statsmodels-developers, 2021a. SARIMAX: Introduction [WWW Document]. statsmodels. URL https://www.statsmodels.org/dev/generated/statsmodels.tsa.statespace.sarimax.SARIMAX.html (accessed 9.5.21).

Perktold, Skipper, S., Taylor, J., statsmodels-developers, 2021b. statsmodels.tsa.statespace.sarimax.SARIMAX — statsmodels [WWW Document]. statsmodels. URL https://www.statsmodels.org/dev/generated/statsmodels.tsa.statespace.sarimax.SARIMAX.html (accessed 9.5.21).

Pham, D.N., Aziz, T., Kohan, A., Nellis, S., Jamil, J. b A., Khoo, J.J., Lukose, D., AbuBakar, S., Sattar, A., Ong, H.H., 2018. How to Efficiently Predict Dengue Incidence in Kuala Lumpur, in: 2018 Fourth International Conference on Advances in Computing, Communication Automation (ICACCA). Presented at the 2018 Fourth International

Conference on Advances in Computing, Communication Automation (ICACCA), pp. 1–6. https://doi.org/10.1109/ICACCAF.2018.8776790

Pham, H.V., Doan, H.T., Phan, T.T., Tran Minh, N.N., 2011. Ecological factors associated with dengue fever in a central highlands Province, Vietnam. BMC Infect. Dis. 11, 172. https://doi.org/10.1186/1471-2334-11-172

Pham, N.T., Nguyen, C.T., Pineda-Cortel, M.R.B., 2020. Time-series modelling of dengue incidence in the Mekong Delta region of Viet Nam using remote sensing data. West. Pac. Surveill. Response J. WPSAR 11, 13–21. https://doi.org/10.5365/wpsar.2018.9.2.012

Phung, Chu, C., Rutherford, S., Nguyen, H.L.T., Luong, M.A., Do, C.M., Huang, C., 2017. Heavy rainfall and risk of infectious intestinal diseases in the most populous city in Vietnam. Sci. Total Environ. 580, 805–812. https://doi.org/10.1016/j.scitotenv.2016.12.027

Phung, Huang, C., Rutherford, S., Chu, C., Wang, X., Nguyen, M., Nguyen, N.H., Do, C.M., Nguyen, T.H., 2015a. Temporal and spatial patterns of diarrhoea in the Mekong Delta area, Vietnam. Epidemiol. Infect. 143, 3488–3497. https://doi.org/10.1017/S0950268815000709

Phung, Huang, C., Rutherford, S., Chu, C., Wang, X., Nguyen, M., Nguyen, N.H., Manh, C.D., 2015b. Identification of the prediction model for dengue incidence in Can Tho city, a Mekong Delta area in Vietnam. Acta Trop. 141, 88–96. https://doi.org/10.1016/j.actatropica.2014.10.005

Phung, Huang, C., Rutherford, S., Chu, C., Wang, X., Nguyen, M., Nguyen, N.H., Manh, C.D., Nguyen, T.H., 2015c. Association between climate factors and diarrhoea in a Mekong Delta area. Int. J. Biometeorol. 59, 1321–1331. https://doi.org/10.1007/s00484-014-0942-1

Phung, Nguyen, H.X., Nguyen, H.L.T., Luong, A.M., Do, C.M., Tran, Q.D., Chu, C., 2018. The effects of socioecological factors on variation of communicable diseases: A multiple-disease study at the national scale of Vietnam. PLOS ONE 13, e0193246. https://doi.org/10.1371/journal.pone.0193246

Phuong, L.T.D., Hanh, T.T.T., Nam, V.S., 2016. Climate Variability and Dengue Hemorrhagic Fever in Ba Tri District, Ben Tre Province, Vietnam during 2004-2014. AIMS Public Health 3, 769–780. https://doi.org/10.3934/publichealth.2016.4.769

Roback, P., Legler, J., 2021. Chapter 4 Poisson Regression | Beyond Multiple Linear Regression: Applied Generalized Linear Models and Multilevel Models in R., 1st ed. Chapman and Hall/CRC Press.

Sahai, A.K., Mandal, R., Joseph, S., Saha, S., Awate, P., Dutta, S., Dey, A., Chattopadhyay, R., Phani, R., Pattanaik, D.R., Despande, S., 2020. Development of a probabilistic early health warning system based on meteorological parameters. Sci. Rep. 10, 14741. https://doi.org/10.1038/s41598-020-71668-6

Schmidhuber, J., 2015. Deep learning in neural networks: An overview. Neural Netw. 61, 85–117. https://doi.org/10.1016/j.neunet.2014.09.003

Scott, T.W., Amerasinghe, P.H., Morrison, A.C., Lorenz, L.H., Clark, G.G., Strickman, D., Kittayapong, P., Edman, J.D., 2000. Longitudinal Studies of Aedes aegypti (Diptera: Culicidae) in Thailand and Puerto Rico: Blood Feeding Frequency. J. Med. Entomol. 37, 89–101. https://doi.org/10.1603/0022-2585-37.1.89

Seabold, S., Perktold, J., 2010. Statsmodels: Econometric and Statistical Modeling with Python. Proc. 9th Python Sci. Conf. 2010.

Shi, Y., Liu, X., Kok, S.-Y., Rajarethinam, J., Liang, S., Yap, G., Chong, C.-S., Lee, K.-S., Tan, S.S.Y., Chin, C.K.Y., Lo, A., Kong, W., Ng, L.C., Cook, A.R., 2016. Three-Month Real-Time Dengue Forecast Models: An Early Warning System for Outbreak Alerts

and Policy Decision Support in Singapore. Environ. Health Perspect. 124, 1369–1375. https://doi.org/10.1289/ehp.1509981

Socialist Republic of Viet Nam, 2010. Viet Nam's Second National Communication to the UNFCCC.

ten Bosch, Q.A., Singh, B.K., Hassan, M.R.A., Chadee, D.D., Michael, E., 2016. The Role of Serotype Interactions and Seasonality in Dengue Model Selection and Control: Insights from a Pattern Matching Approach. PLoS Negl. Trop. Dis. 10. https://doi.org/10.1371/journal.pntd.0004680

Terpilowski, M.A., 2019. scikit-posthocs: Pairwise multiple comparison tests in Python. J. Open Source Softw. 4, 1169. https://doi.org/10.21105/joss.01169

Thompson, C.N., Phan, M.V.T., Hoang, N.V.M., Minh, P.V., Vinh, N.T., Thuy, C.T., Nga, T.T.T., Rabaa, M.A., Duy, P.T., Dung, T.T.N., Phat, V.V., Nga, T.V.T., Tu, L.T.P., Tuyen, H.T., Yoshihara, K., Jenkins, C., Duong, V.T., Phuc, H.L., Tuyet, P.T.N., Ngoc, N.M., Vinh, H., Chinh, N.T., Thuong, T.C., Tuan, H.M., Hien, T.T., Campbell, J.I., Chau, N.V.V., Thwaites, G., Baker, S., 2015a. A Prospective Multi-Center Observational Study of Children Hospitalized with Diarrhea in Ho Chi Minh City, Vietnam. Am. J. Trop. Med. Hyg. 92, 1045–1052. https://doi.org/10.4269/ajtmh.14-0655

Thompson, C.N., Zelner, J.L., Nhu, T.D.H., Phan, M.V., Hoang Le, P., Nguyen Thanh, H., Vu Thuy, D., Minh Nguyen, N., Ha Manh, T., Van Hoang Minh, T., Lu Lan, V., Nguyen Van Vinh, C., Tran Tinh, H., von Clemm, E., Storch, H., Thwaites, G., Grenfell, B.T., Baker, S., 2015b. The impact of environmental and climatic variation on the spatiotemporal trends of hospitalized pediatric diarrhea in Ho Chi Minh City, Vietnam. Health Place 35, 147–154. https://doi.org/10.1016/j.healthplace.2015.08.001

Tjaden, N.B., Thomas, S.M., Fischer, D., Beierkuhnlein, C., 2013. Extrinsic Incubation Period of Dengue: Knowledge, Backlog, and Applications of Temperature Dependence. PLoS Negl. Trop. Dis. 7, e2207. https://doi.org/10.1371/journal.pntd.0002207

Troeger, C., Blacker, B.F., Khalil, I.A., Rao, P.C., Cao, S., Zimsen, S.R., Albertson, S.B., Stanaway, J.D., Deshpande, A., Abebe, Z., Alvis-Guzman, N., Amare, A.T., Asgedom, S.W., Anteneh, Z.A., Antonio, C.A.T., Aremu, O., Asfaw, E.T., Atey, T.M., Atique, S., Avokpaho, E.F.G.A., Awasthi, A., Ayele, H.T., Barac, A., Barreto, M.L., Bassat, Q., Belay, S.A., Bensenor, I.M., Bhutta, Z.A., Bijani, A., Bizuneh, H., Castañeda-Orjuela, C.A., Dadi, A.F., Dandona, L., Dandona, R., Do, H.P., Dubey, M., Dubljanin, E., Edessa, D., Endries, A.Y., Eshrati, B., Farag, T., Feyissa, G.T., Foreman, K.J., Forouzanfar, M.H., Fullman, N., Gething, P.W., Gishu, M.D., Godwin, W.W., Gugnani, H.C., Gupta, R., Hailu, G.B., Hassen, H.Y., Hibstu, D.T., Ilesanmi, O.S., Jonas, J.B., Kahsay, A., Kang, G., Kasaeian, A., Khader, Y.S., Khalil, I.A., Khan, E.A., Khan, M.A., Khang, Y.-H., Kissoon, N., Kochhar, S., Kotloff, K.L., Koyanagi, A., Kumar, G.A., Magdy Abd El Razek, H., Malekzadeh, R., Malta, D.C., Mehata, S., Mendoza, W., Mengistu, D.T., Menota, B.G., Mezgebe, H.B., Mlashu, F.W., Murthy, S., Naik, G.A., Nguyen, C.T., Nguyen, T.H., Ningrum, D.N.A., Ogbo, F.A., Olagunju, A.T., Paudel, D., Platts-Mills, J.A., Qorbani, M., Rafay, A., Rai, R.K., Rana, S.M., Ranabhat, C.L., Rasella, D., Ray, S.E., Reis, C., Renzaho, A.M., Rezai, M.S., Ruhago, G.M., Safiri, S., Salomon, J.A., Sanabria, J.R., Sartorius, B., Sawhney, M., Sepanlou, S.G., Shigematsu, M., Sisay, M., Somayaji, R., Sreeramareddy, C.T., Sykes, B.L., Taffere, G.R., Topor-Madry, R., Tran, B.X., Tuem, K.B., Ukwaja, K.N., Vollset, S.E., Walson, J.L., Weaver, M.R., Weldegwergs, K.G., Werdecker, A., Workicho, A., Yenesew, M., Yirsaw, B.D., Yonemoto, N., El Sayed Zaki, M., Vos, T., Lim, S.S., Naghavi, M., Murray, C.J., Mokdad, A.H., Hay, S.I., Reiner, 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. Lancet Infect. Dis. 18, 1211–1228. https://doi.org/10.1016/S1473-3099(18)30362-1

Tuyet-Hanh, T.T., Ha, V.N., Hoang, V.M., Le, T.T.H., Nguyen, T.T.N., Tran, M.K., Nguyen, H.Q., Luu, Q.T., Tran, N.Q.L., Tran, V.A., 2018a. Health Vulnerability and Adaptation Assessment in Viet Nam. Climate Change Research Group, Hanoi University of Public Health, Hanoi.

Tuyet-Hanh, T.T., Nhat Cam, N., Thi Thanh Huong, L., Khanh Long, T., Mai Kien, T., Thi Kim Hanh, D., Huu Quyen, N., Nu Quy Linh, T., Rocklöv, J., Quam, M., Van Minh, H., 2018b. Climate Variability and Dengue Hemorrhagic Fever in Hanoi, Viet Nam, During 2008 to 2015. Asia Pac. J. Public Health 30, 532–541. https://doi.org/10.1177/1010539518790143

UNICEF, 2018. Children in Viet Nam [WWW Document]. UNICEF. URL https://www.unicef.org/vietnam/children-viet-nam (accessed 8.18.21).

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is All you Need, in: Advances in Neural Information Processing Systems 30. pp. 5998–6008.

Villabona-Arenas, C.J., Oliveira, J.L. de, Capra, C. de S., Balarini, K., Loureiro, M., Fonseca, C.R.T.P., Passos, S.D., Zanotto, P.M. de A., 2014. Detection Of Four Dengue Serotypes Suggests Rise In Hyperendemicity In Urban Centers Of Brazil. PLoS Negl. Trop. Dis. 8, e2620. https://doi.org/10.1371/journal.pntd.0002620

Virtanen, P., Gommers, R., Oliphant, T.E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S.J., Brett, M., Wilson, J., Millman, K.J., Mayorov, N., Nelson, A.R.J., Jones, E., Kern, R., Larson, E., Carey, C.J., Polat, İ., Feng, Y., Moore, E.W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E.A., Harris, C.R., Archibald, A.M., Ribeiro, A.H., Pedregosa, F., van Mulbregt, P., 2020. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat. Methods 17, 261–272. https://doi.org/10.1038/s41592-019-0686-2

Wang, C., Jiang, B., Fan, J., Wang, F., Liu, Q., 2014. A Study of the Dengue Epidemic and Meteorological Factors in Guangzhou, China, by Using a Zero-Inflated Poisson Regression Model. Asia Pac. J. Public Health 26, 48–57. https://doi.org/10.1177/1010539513490195

Wangdi, K., Clements, A.C., 2017. Spatial and temporal patterns of diarrhoea in Bhutan 2003–2013. BMC Infect. Dis. 17, 507. https://doi.org/10.1186/s12879-017-2611-6

Waskom, M.L., 2021. seaborn: statistical data visualization. J. Open Source Softw. 6, 3021. https://doi.org/10.21105/joss.03021

Wilke, A.B.B., Chase, C., Vasquez, C., Carvajal, A., Medina, J., Petrie, W.D., Beier, J.C., 2019. Urbanization creates diverse aquatic habitats for immature mosquitoes in urban areas. Sci. Rep. 9, 15335. https://doi.org/10.1038/s41598-019-51787-5

World Health Organisation, 2020. Dengue and severe dengue [WWW Document]. World Health Organ. URL https://www.who.int/news-room/fact-sheets/detail/dengue-and-severe-dengue (accessed 1.17.21).

World Health Organisation, 2017. Diarrhoeal disease [WWW Document]. URL https://www.who.int/news-room/fact-sheets/detail/diarrhoeal-disease (accessed 8.18.21).

Xu, J., Xu, K., Li, Z., Meng, F., Tu, T., Xu, L., Liu, Q., 2020. Forecast of Dengue Cases in 20 Chinese Cities Based on the Deep Learning Method. Int. J. Environ. Res. Public. Health 17, 453. https://doi.org/10.3390/ijerph17020453

Zeyer, A., Bahar, P., Irie, K., Schlüter, R., Ney, H., 2019. A Comparison of Transformer and LSTM Encoder Decoder Models for ASR, in: 2019 IEEE Automatic Speech

Recognition and Understanding Workshop (ASRU). pp. 8–15. https://doi.org/10.1109/ASRU46091.2019.9004025

Zhao, Y., Zhu, Y., Zhu, Z., Qu, B., 2016. Association between meteorological factors and bacillary dysentery incidence in Chaoyang city, China: an ecological study. BMJ Open 6, e013376. https://doi.org/10.1136/bmjopen-2016-013376

Zhu, X., Fu, B., Yang, Y., Ma, Y., Hao, J., Chen, S., Liu, Shuang, Li, T., Liu, Sen, Guo, W., Liao, Z., 2019. Attention-based recurrent neural network for influenza epidemic prediction. BMC Bioinformatics 20, 575. https://doi.org/10.1186/s12859-019-3131-8