UTS UNIVERSITY OF TECHNOLOGY SYDNEY

Modelling the Transmission of Dengue Fever Based on Spatial and Temporal Patterns

by Ali Hasan M Siddiq

Thesis submitted in fulfilment of the requirements for the degree of

Doctor of Philosophy

under the supervision of Dr Nagesh Shukla and Distinguished Professor Biswajeet Pradhan

University of Technology Sydney
Faculty of Engineering and Information Technology

February 2023

CERTIFICATE OF ORIGINAL AUTHORSHIP/ORIGINALITY

I, ***Ali Hasan M Siddiq*** declare that this thesis is submitted in fulfilment of the requirements for the award of **Doctor of Philosophy**, in the **School of Professional Practice and Leadership**, **Faculty of Engineering and Information Technology FEIT** at the **University of Technology Sydney**.

This thesis is wholly my own work unless otherwise referenced or acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

This document has not been submitted for qualifications at any other academic institution.

This research is supported by the Australian Government Research Training Program.

Signature:

Production Note:
Signature removed prior to publication.

Ali Hasan M Siddiq

Date: 20/02/2023

COPYRIGHT

DEDICATION

I dedicate this to my Lord and Saviour, in thanks for giving me a talent and setting me on the course to discover and develop it.

To my family and friends, I appreciate your support and encouragement during my PhD journey.

This thesis is dedicated to the greatest person in the world: my brilliant, supportive and loving father, **Hasan Siddiq**, who has always encouraged me to open new doors, challenge myself, and to uphold the principles of integrity at all times. I am honoured to carry your name and am privileged to have you as a father. Many thanks go to my wonderful mother, **Najah Alawaji,** whose life is a reflection of diligence and strength, and all the **Siddiq** family who gave me support through this endeavour. I thank my loving wife **Soha Alsam**, who has endured four years of my working toward this degree. Only through her encouragement, faith and support have I been able to be successful and be where I am today. I am grateful to have two beautiful daughters, **Rayana** and **Joanna** who continue to make my life and work a joy each day. All of you are, and always will be, the source of my strength and determination.

To my friends both inside and outside of academia, thanks for keeping me sane and grounded. It's easy to forget how to have fun during grad school, but you all ensured that I never fell into that trap.

The hard work and sacrifices of our grandparents, parents, aunts, uncles, immediate and extended family can never be repaid but will always be remembered and cherished. The love of our families continues to, and will always, drive us to be better.

Once again, thank you for making this possible and I would like to dedicate this thesis to ALL OF YOU.

ACKNOWLEDGMENTS

A PhD is a long and arduous journey, and now that I have come to the end of my journey. I've been lucky to have been surrounded by a truly wonderful group of people over the last three and a half years, and I know for a fact that I wouldn't have made it this far without them.

A work like this can be achieved only with the help and support of a large group of people and institutions. This section may be extensive and tedious but since the people who actually bother to read it, will look for their names in it I am going to make an effort to not omit anyone. Thus, I would like to extend tremendous thanks and appreciation to several people for their consistent encouragement and guidance throughout this entire experience. The credit is as much yours as it is my own.

First and foremost, I would like to express my deepest gratitude to my supervisor Dr Nagesh Shukla for his invaluable guidance and continuous support throughout my years at UTS, and for giving me the opportunity to work with him and learn from him. He never forced an agenda on my research but instead consistently guided me to pursue my own path, and gave me both the support and the freedom to carry out my research. I would also like to extend my thanks to my supervisor Distinguished Professor Dr Biswajeet Pradhan who has been extremely patient and taught me a great deal, unstintingly sharing his creativity and wealth of knowledge, and introducing me to new research methods and ideas. These experiences have been crucial in shaping how I have approached dengue fever spatiotemporal modelling questions in this thesis. Without my supervisors' leadership and insights, and despite their extremely busy schedules, they have always made themselves available for discussions and guidance. I wanted to reiterate my thanks; without their help, none of this research would have been possible. I look forward to lifelong personal and professional relationships with them.

I'd also be remiss to not mention and extend my lasting gratitude to everyone who had some substantial input in the project; special thanks to Dr Mohsen Naderpour, Dr Babak Abedin, Dr David Milne and all of the candidature assessment panel members for their discussions and suggestions. This thesis greatly benefited from the insights, constructive criticism, and many helpful suggestions provided by the PhD committee.

I would like to acknowledge the support of the Saudi Ministry of Health for their approval to conduct this research and obtain the data related to the study. Thanks also to Dr. Ali Alzahrani (Director) and Dr. Hayel Qudsi (Vice Director) of the Vector-borne & Zoonotic Diseases Administration of Jeddah for their generosity in making available the records of reported cases of dengue and providing the data pertaining to reported dengue cases in Jeddah from 2012 to 2018.

I wish to express my gratitude to both the Saudi government and to my employer, Jazan University. The scholarship they awarded me, which included living expenses and fees, made my life in Sydney and UTS possible. Thank you for believing in me, giving me this opportunity to study for my PhD degree, and making it financially possible.

My PhD journey was a struggle at times, and I appreciate all the assistance and encouragement that has been given to me. This thesis is as much yours as it is mine.

We started together, and now we have finally finished.

Thanks to all for getting me through it.

LIST OF PAPERS/PUBLICATIONS

Published journal articles

1. A. Siddiq, N. Shukla, and B. Pradhan, "Spatio-temporal Modelling of Dengue Fever Cases in Saudi Arabia using Socio-economic, Climatic and Environmental Factors," Geocarto International, pp. 1-23, 2022, doi: 10.1080/10106049.2022.2072005.

Submitted journal articles

1. Modelling the Transmission of Dengue Fever Based on Spatial and Temporal Patterns: Systematic Review
2. Dengue fever prediction modelling considering missing data imputations
3. Spatiotemporal Simulation of Dengue Fever Cases in Jeddah City using Cellular Automata Model

Published conference papers

1. Siddiq, A., Shukla, N. & Pradhan, B. 2021, 'Predicting Dengue Fever Transmission Using Machine Learning Methods', IEEE, pp. 21-6.

All of the aforementioned papers have been published during my Ph.D. candidature.

Abstract of thesis presented to the Senate of University of Technology Sydney in fulfilment of the requirement for the degree of Doctor of Philosophy

PUBLICATIONS INCLUDED IN THIS THESIS

Publication citation – incorporated within chapters in a conventional form.

Modelling the Transmission of Dengue Fever Based on Spatial and Temporal Patterns: Systematic Review

Contributor	Statement of contribution	Thesis chapters	Status
Ali Siddiq	**A.S** initiated the study topic and searched the data bases for relevant literatures. A.S also performed the literature review analysis and developed the manuscript while following guidance from (**N.S and B.P**) supervisors.	Chapter 2	Under review
Nagesh Shukla	**N.S** guided the literature review analysis, review and approved the manuscript.		
Biswajeet Pradhan	**B.P** guided the literature review analysis, review and approved the manuscript.		

A. Siddiq, N. Shukla, and B. Pradhan, "Spatio-temporal Modelling of Dengue Fever Cases in Saudi Arabia using Socio-economic, Climatic and Environmental Factors," Geocarto International, pp. 1-23, 2022, doi: 10.1080/10106049.2022.2072005.

Contributor	Statement of contribution	Thesis chapters	Status
Ali Siddiq	**A.S** collected research related datasets. A.S also performed the data analysis and developed the manuscript while following guidance from (**N.S and B.P**) supervisors.	Chapters 1, 3, 4, and 5	Published
Nagesh Shukla	**N.S** guided the literature review analysis, review and approved the manuscript.		
Biswajeet Pradhan	**B.P** guided the literature review analysis, review and approved the manuscript.		

Dengue fever prediction modelling considering missing data imputations

Contributor	Statement of contribution	Thesis chapters	Status
Ali Siddiq	**A.S** collected research related datasets. A.S also performed the data analysis and developed the manuscript while following guidance from (**N.S and B.P**) supervisors.	Chapters 1, 3, 4, and 5	Under review
Nagesh Shukla	**N.S** guided the literature review analysis, review and approved the manuscript.		
Biswajeet Pradhan	**B.P** guided the literature review analysis, review and approved the manuscript.		

Spatiotemporal Simulation of Dengue Fever Cases in Jeddah City using Cellular Automata Model

Contributor	Statement of contribution	Thesis chapters	Status
Ali Siddiq	**A.S** collected research related datasets. A.S also performed the data analysis and developed the manuscript while following guidance from (**N.S and B.P**) supervisors.	Chapters 1, 3, 4, and 5	Under review
Nagesh Shukla	**N.S** guided the literature review analysis, review and approved the manuscript.		
Biswajeet Pradhan	**B.P** guided the literature review analysis, review and approved the manuscript.		

Siddiq, A., Shukla, N. & Pradhan, B. 2021, 'Predicting Dengue Fever Transmission Using Machine Learning Methods', IEEE, pp. 21-6

Contributor	Statement of contribution	Thesis chapters	Status
Ali Siddiq	**A.S** collected research related datasets. A.S also performed the data analysis and developed the manuscript while following guidance from (**N.S and B.P**) supervisors.	Chapters 1, 3, 4, and 5	Published
Nagesh Shukla	**N.S** guided the literature review analysis, review and approved the manuscript.		
Biswajeet Pradhan	**B.P** guided the literature review analysis, review and approved the manuscript.		

To those who made all of this possible…

It is finished, finally.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AIC	Akaike Information Criterion
ANOVA	Analysis of Variance
AUC	Area Under Curve
AI	Artificial Intelligence
ANN	Artificial Neural Network
BME	Bayesian Maximum Entropy
BRT	Boosted Regression Tree
BI	Breteau Index
CA	Cellular Automata
CDC	Centres for Disease Control and Prevention
CART	Classification and Regression Tree
CCI	Climate Change Initiative
CHIRPS	Climate Hazards Group InfraRed Precipitation with Station
CAR	Conditional Autoregressive
CI	Container Index
DT	Decision Tree
DTR	Decision Trees Regression
DF	Dengue Fever
DHF	Dengue Haemorrhagic Fever
DSS	Dengue Shock Syndrome
DENV	Dengue Virus Serotype
DENV1	Dengue Virus Type 1
DENV2	Dengue Virus Type 2
DENV3	Dengue Virus Type 3
DENV4	Dengue Virus Type 4
DBSCAN	Density-Based Spatial Clustering of Applications with Noise
DIC	Deviance Information Criterion
DEM	Digital Elevation Model
EBS	Empirical Bayes Smoothing
ESTARFM	Enhanced Spatial and Temporal Adaptive Reflectance Fusion Model
EVI	Enhanced Vegetation Index

ESA	European Space Agency
FARM	Fuzzy Association Rule Mining
GAMMs	Generalized Additive Mixed Models
GAM	Generalized Additive Models
GLMM	Generalized Linear Mixed Models
GLM	Generalized Linear Model
GARP	Genetic Algorithm for Rule-Set Prediction
GIS	Geographical Information System
GWPR	Geographically Weighted Poisson Regression
GWR	Geographically Weighted Regression
GEE	Google Earth Engine
GDP	Gross Domestic Product
HI	House Index
HIF	House Infestation Index
ISODATA	Iterative Self-Organizing Data Analysis Techniques Algorithm
KDE	Kernel-Density Estimation
KSA	Kingdom of Saudi Arabia
KNN	K-Nearest Neighbour
LPDAAC	Land Processes Distributed Active Archive Center
LST	Land Surface Temperature
LUL	Land Urbanization Level
LULC	Land Use/Land Cover
LISA	Local Indicators of Spatial Association
LSAS	Local Spatial Autocorrelation Statistic
ML	Machine Learning
MCMC	Markov Chain Monte Carlo
MOH	Ministry of Health
MAR	Missing at Random
MCAR	Missing Completely at Random
MD	Missing Data
MNAR	Missing Not at Random
DI	Density Index
MLP	Multilayer Perceptron
MI	Multiple Imputation

NCAR	National Center for Atmospheric Research
NCEP	National Centres for Environmental Information
NOAA	National Oceanic and Atmospheric Administration
NPV	Negative Predictive Value
NDVI	Normalized Difference Vegetation Index
NDWI	Normalized Difference Water Index
ODE	Ordinary Differential Equation
OLS	Ordinary Least Square
PRISM	Parameter-Elevation Regressions on Independent Slopes Model
PPA	Point Pattern Analysis
PPV	Positive Predictive Value
PCA	Principal Component Analysis
RF	Random Forest
ROC	Receiver Operating Characteristics
RS	Remote Sensing
RMSE	Root Mean Squared Error
SST	Sea Surface Temperature
SOFM	Self Organizing Feature Map
SOM	Self-Organized Map
SOI	Southern Oscillation Index
SAC	Spatial Autocorrelation
SAR	Spatial Autoregressive Specification
SDM	Species Distribution Modelling
SDE	Standard Deviational Ellipsis
SRMSE	Standard Root Mean Squared Error
SVC	Support Vector Classification
SVM	Support Vector Machines
USGS	US Geological Survey
VIF	Variance Inflation Factor
VBD	Vector-Borne Disease
WHO	World Health Organization

ABSTRACT

Dengue fever (DF) is a vector-borne disease that has transmit alarmingly in recent decades and has now affected the populations of roughly 100 nations, primarily in tropical and subtropical regions. Approximately 390 million cases of dengue fever are reported each year among people living in 128 countries, according to the World Health Organization (WHO). Viral, host, and vector interactions result in complex spatiotemporal patterns in dengue disease. Moreover, it has been previously indicated that the dengue fever epidemic is due to several climatic, social, environmental, and biological factors, and these factors vary from place to place and with time. Therefore, an accurate spatiotemporal prediction model is essential to understand the disease patterns and improve the monitoring and control of potential threats.

Several research gaps in previous DF spatiotemporal prediction models are addressed in this work: (i) the lack of a comprehensive framework that ensures better spatiotemporal prediction models; (ii) the lack of work on missing spatiotemporal data and data quality; (iii) the lack of testing and comparison between the performance of different traditional and advanced spatiotemporal modelling approaches; and (iv) the lack of consideration of simulation maps based on optimal pre-processing analysis as a means of controlling future disease threats. This research is intended to address these shortcomings and contribute to the literature by: (i) improving the current understanding of the spatiotemporal patterns of DF; (ii) developing a comprehensive framework to improve the spatial and temporal prediction models' performance, and can effectively estimate the potential risk of disease that will help authorities allocate resources and implement effective control measures at a district scale; (iii) examining previous DF spatiotemporal modelling approaches and significant disease-related factors using a

geographical information system and machine learning methods; (iv) analysing and improving the quality of collected data and investigating several traditional and advanced imputation approaches used to fill missing values in order to provide more reliable and unbiased results; and (v) visualising a simulation of potential future risk areas as a simple mechanism to assist decision-makers control the disease.

To achieve the first objective, after reviewing the previous spatiotemporal studies and analysing the adopted models and considered factors; common factors were collected to be used in the proposed framework. Subsequently, the adopted methods were integrated and applied to investigate the spatiotemporal patterns of the disease. First, Geographical Information Systems (GIS) and spatial statistics were applied to improve the current understanding of dengue patterns. Moreover, spatial analysis was conducted using the ordinary least square (OLS) and geographically weighted regression (GWR) models to identify high-risk areas. Then, a robust data analysis model was developed to provide a better understanding of confirmed dengue fever cases despite missing data (MD). In addition, better insights were obtained on the risk factors associated with confirmed cases. Lastly, machine learning algorithms were utilized to create clusters of patients with comparable characteristics.

For the second objective, this study investigated the importance of extensive analysis required to increase the prediction accuracy of disease-confirmed cases. It was noted that the accuracy of the DF spatiotemporal model depends on the quality of the observed data (such as MD in the input). Thus, it was necessary to create data models that can deal effectively with the imputation of MD, as failure to do so is likely to produce erroneous results. A cluster-based technique was applied to determine the appropriate values to replace any missing values. Utilizing the results obtained from several scenarios, the performance of the proposed approach was compared with that of several traditional

and complex imputation algorithms. The proposed model significantly improved the accuracy of the prediction model, demonstrating that the proposed imputation method outperforms traditional methods.

In order to achieve the third objective, simulation maps were created to quantify districts at high risk due to impending dengue disease. The cellular automata approach was used to simulate risk maps at a district scale. Subsequently, a validation process was conducted to compare simulation maps with the observed data. The results indicate that the proposed model can predict future cases significantly. These results suggest that the simulation is a valuable means of obtaining a better understanding of spatiotemporal complexity in dynamic spatial phenomena, such as dengue fever. The model was applied to the city of Jeddah in the Kingdom of Saudi Arabia, as it recorded the highest number of DF cases in the country. The research findings can assist policymakers to develop effective methods and control procedures against potential DF threats.

The significance of the current work is its transferability of the proposed framework to other regions, which allows health authorities to optimize prevention planning by localizing the input parameters that contribute to DF spatial and spatiotemporal transmission. Moreover, health authorities can use existing data locally to predict epidemics in the near future, so that measures can be taken to prevent new cases from occurring and prepare local governments for potential crises. Thus, a precise risk map can also be used to allocate resources for effective district-level disease spatial and spatiotemporal transmission control. Lastly, the current thesis contributes to researchers by offering a comprehensive framework that involves several analysis stages to improve the spatiotemporal model performance. As a result, it will encourage researchers to explore other methods of improving a model's accuracy.

Keywords: Dengue Fever (DF), Geographical Information System (GIS), Self-Organized Map (SOM), Density-Based Spatial Clustering of Applications with Noise (DBSCAN), Spatiotemporal Modelling, Machine Learning, Cellular Automata, Simulation, Prediction Model.

CHAPTER 1

INTRODUCTION

This chapter presents a general introduction to the research topic, the research background and previous similar modelling researches. This chapter introduces the significant factors that have been found to influence the transmission of DF. This chapter also outlines the primary purpose of the study, formulates the problem statement and explains the motivation for this study followed by the research questions, discusses research gaps as well as the research hypothesis, states the research aims and objectives and scope, the significance and novelty of the research are explained and sets out the thesis organization. It addresses the impacts of the disease and discusses the importance of the proposed modelling framework to stakeholders including those authorities tasked with controlling the disease.

1.1 General introduction

In its 2017 report, the World Health Organization (WHO) defined vector-borne diseases (VBDs) as "human illnesses caused by parasites, viruses and bacteria that are transmitted by vectors" (Taghikhani 2020; World Health Organization 2017). In this report, the WHO also stated that, annually, 700,000 deaths around the globe are caused by VBDs. Of this number, malaria kills over 400,000 people worldwide each year, and hundreds of millions of people worldwide have been affected by other diseases such as Chagas, Leishmaniasis and Schistosomiasis. Additionally, around four billion people across more than 128 nations are in danger of contracting dengue fever (DF) (Chen et al. 2018; Huang et al. 2018; Vincenti-Gonzalez et al. 2017; Whiteman et al. 2019). The annual number of worldwide symptomatic cases of DF is in excess of 96 million, resulting

in an estimated 40,000 deaths (World Health Organization 2019). In fact, after malaria, DF is considered to be the second most severe infectious and life-threatening VBD (Badreddine et al. 2017).

DF is a mosquito-borne disease that is transmitted by numerous types of mosquitoes from the *Aedes* family of *Flaviviridae* (*Aedes Aegypti "Ae. aegypti", Aedes Albopictus "Ae. albopictus", and Aedes Polynesiensis* "tiger mosquito") (Astuti et al. 2019; Dhewantara et al. 2019; Jeefoo 2012; Vincenti-Gonzalez et al. 2017). *Ae. aegypti* tends to be the main vector because of its proximity to human residences (Altassan et al. 2019; Organji et al. 2017). Moreover, the female of *Ae. aegypti* is viewed as the essential vector of DF (Abou El-Saoud et al. 2018; Khormi et al. 2011). There are four antigenically different, but associated, dengue virus serotypes identified as DENV-1 to DENV-4 (Jeefoo 2012; Nguyen et al. 2020; Rotela et al. 2007; Yung et al. 2015); each serotype can cause a range of the disease's symptoms (Gubler 1998). A minority of patients acquire a severe form of the disease, which starts with DF and has an uneven development of severe vascular leakage, which can quickly result in shock or death; this is known as dengue haemorrhagic fever (DHF) / dengue shock syndrome (DSS) (Gibbons and Vaughn 2002; Khan et al. 2008). A fifth serotype has been identified recently, but little is known about it (Altassan et al. 2019).

Several studies have been conducted worldwide in order to better understand the spatial and spatiotemporal patterns and factors that influence the transmission of DF, as discussed in more detail in the second chapter. Moreover, understanding the interplay between DF and possible risk factors (climatic, demographic and socio-economic, entomological, and environmental), which are connected to the spatiotemporal patterns of disease, is critical, and will assist authorities in taking appropriate control measures (Cao et al. 2017; Teurlai et al. 2015). Therefore, this study aims to develop a novel

framework for producing high-performance dengue spatiotemporal models on a district scale.

1.2 Research background

Main factors contributing to DF disease transmission in previous studies

Many variables have contributed to the worldwide increase of DF over the previous 50 years, some of which are: increased urbanisation, migration of local and foreign communities, erratic water supplies and the geographically increasing climate-related vectors (Altassan et al. 2019). Moreover, in addition to the urbanisation factor, the other principal factors that have driven the emergence of epidemic dengue are globalisation and the absence of efficient mosquito control (Gubler 2011; Kholedi et al. 2012). International travel is also a significant means by which DF is transmitting from nation to nation. Previous studies have shown that imported dengue instances can initiate native epidemics if the climate is suitable (Wen and Tsai 2016). In some nations, humans are living in crowded conditions close to massive numbers of mosquitoes. This dense population encourages the transmission of the dengue virus between people due to the vector's "mosquito-human transmission cycle" (Gubler 2011). Moreover, the spatial transmission of DF is feasible only when either one or both of the infected organism pathogens move (Enduri and Jolad 2018). The *Ae. aegypti* is found mostly around urban environments, in man-made conduits and containers, and in stagnant water (Valles et al. 2019). Moreover, the incidence of DF is correlated in many places with vegetation indexes, tree cover and soil cover, since these habitat features affect the size of the vector population. Additionally, in regions with buildings and medium-sized trees, *Ae. aegypti* are more likely to be present (Altassan et al. 2019). Generally, almost 75% of all infection diseases contracted by humans comes from animals (Al-Tayib 2019). Moreover, this large number

is an indication of newly-emerging viruses over the last two decades. Therefore, the people working and living close to an animal-populated environment are considered to be potential factors in the transmission of the disease (Al-Tayib 2019). While the real impact of environmental and climatic changes on vector abundance and illnesses has yet to be confirmed, the effect on dengue vectors of environmental modifications resulting from human behaviours and social changes due to climate change is likely (Higa 2011). There are still significant issues when applying dengue models to real-world decision issues, due to the limited knowledge of the variables affecting the transmission of dengue and the restricted accessibility to information (Andraud et al. 2012). Hence, the initial steps toward controlling, monitoring, or predicting the transmission of DF involves determining and understanding the main variables that have the potential to have an impact on disease transmission.

The impact of data quality on DF modelling performance accuracy

Missing data (MD) refers to the absence of values for factors in established parameters (Kang 2013; Ngueilbaye et al. 2021). Despite efforts by researchers to obtain complete data (Scheffer 2002), almost all studies have MD, even those that are well-designed and controlled (Kang 2013). MD is a severe issue in data analysis and decision-making since the obtained information is incomplete, and therefore may be untrustworthy (Ngueilbaye et al. 2021). Moreover, MD can have a detrimental effect on the study's findings (Kang 2013), as it might lead to biased values regarding the target parameters (Kang 2013). Also, during the knowledge discovery process, the risk of having MD increases with the amount of data collected, thereby exacerbating the problem (Brown and Kros 2003). Although many studies do not specifically state how they manage MD (Zhang 2016), a few have focused on ways to deal with it, the difficulties that MD might

create, and how to avoid or reduce them (Kang 2013). The issue is how to deal with MD once it has been determined that the recovery of the actual missing values is impossible (Scheffer 2002). Hence, the problem of data quality has become a recent topic requiring research and investigation (Brown and Kros 2003). Therefore, this aspect was investigated through the previous literature and the methods used to improve the quality of collected data to fit the final model and enhance its performance.

DF modelling, simulation and explanation

The DF spatial patterns are impacted by a complex combination of climatic, environmental, and socio-economic characteristics (Acharya et al. 2018; Delmelle et al. 2016; McGough et al. 2021), which provide significant challenges to dengue transmission studies (De Lima et al. 2016). Additionally, the primary factors for modelling the transmit of the disease vary according to spatial and temporal conditions (Ferrell and Brinkerhoff 2018). Therefore, modelling DF transmission is an appropriate method for better understanding and controlling the disease (Jácome et al. 2019). Moreover, simulations may directly depict the dynamic nature of illness transmission and study the potential causes of the observed effects in a world that is becoming more complicated (Halloran et al. 2017). Thus, understanding the developed model mechanisms and the reasons behind the prediction model is fundamental to gain the decision-maker's trust (Ribeiro et al. 2016). Additionally, these approaches are appropriate to enhance the current understanding of disease spatial and spatiotemporal dynamicity and testing/modifying control procedures before real implementation (Lemos et al. 2017). Lastly, map visualisation is a simple approach to detect the disease patterns in a specific region, and simplify the illustration of forecasting disease transmission patterns in the future for better controlling strategies (Eosina et al. 2016).

1.3 Problem statement

The World Health Organization (WHO) has designated DF as one of the most dangerous vector-borne infectious illnesses in tropical and subtropical regions (Yu et al. 2011). Moreover, DF is considered to be the second most severe infection and life-threatening VBD after malaria (Badreddine et al. 2017). The spatiotemporal patterns of dengue disease are the consequence of complex interactions between the viral, the host, and the vector (Teurlai et al. 2015). Moreover, such interaction together with the weight of each component in the development and transmission of DF epidemics may vary by nation, depending on the unique climate conditions, and cultural and socio-economic context in which the virus circulates (Reiter 2001; Teurlai et al. 2015). Understanding the significant variables contributing to the spatiotemporal patterns of infectious disease can help with disease prevention and control (Dhewantara et al. 2019). Previous studies focus specifically on assessing the related factors and generating risk maps without mentioning the preprocessing stages and their impact on the performance of the proposed model. The quality of collected data is essential to the model performance (Kamkhad et al. 2016). Moreover, the performance analysis and outcomes need to be explained to decision-makers and stakeholders to ensure their trust in the model (Ribeiro et al. 2016). Therefore, this thesis proposes a comprehensive framework that can accurately model the spatiotemporal patterns of DF. It is hoped that the framework will provide an effective approach for other researchers involved in analysing different patterns. In the literature review presented in Chapter 2, previous studies focused on the spatiotemporal modelling of DF, are examined in order to determine various perspectives and findings which will help achieve the research objectives of the current study.

1.4 Research motivation

Worldwide, DF incidence has increased dramatically over the past few decades, putting roughly half of the world's population at risk (World Health Organization 2019). DF cannot be explicitly treated; however, early detection can control disease transmission. DF is a complex disease due to the interaction between various social, economic and environmental factors in a specific area of interest. Moreover, there are still significant issues when applying dengue models to real-world decision issues, due to limited knowledge of the variables affecting the transmission of dengue and the lack of, or restricted access to, information in some regions (Andraud et al. 2012). In addition to the difficulty of obtaining data, the quality of that data is a significant issue when developing a highly accurate prediction model, particularly as an effective model will enable health authorities to control the disease and possibly eliminate future threats.

In this thesis, a comprehensive framework that is appropriate and accurate is proposed to model DF from a spatiotemporal perspective. It is hoped that this proposed framework will be adopted by stakeholders to implement better controlling procedures and to create a sustainable, healthier environment for the human population. In order to reduce the possible disease future threats, this study would contribute to identifying the risk areas through the observed data. Thus, it will provide a better insight to health authorities in order to improve current controlling strategies. Lastly, the thesis findings will be a significant avenue for further work to investigate other related aspects and to improve the model's accuracy.

The city of Jeddah was chosen as the area of interest in this thesis because of its geographical location as the gateway to Mecca for Muslims worldwide. Annually, millions of Muslim people from both inside and outside the country travel to the holy city of Mecca for the *Hajj* and the *Umrah*. Therefore, there is a strong likelihood of DF

transmission during these mass religious gatherings, during *Al-Hajj* and *Al-Umrah*. This study area is therefore significant to the Saudi Ministry of Health, which is responsible for decision-making and policies, as the proposed framework may assist in predicting the potential of DF in the country. It can also assist the world's health organisations as it offers an improved prediction model. Specifically, it will benefit both the Muslims who are fulfilling their religious obligations, and the non-Muslims in the home countries of the pilgrims when the latter return from Mecca.

1.5 Research questions

The overall goal of this research is to develop a comprehensive framework to achieve the most accurate model for DF spatiotemporal patterns in order to ensure that public health policies are implemented effectively. This investigation is focused on understanding the integration of different climatic, socio-economic, and environmental factors that interact to influence dengue disease. To develop a highly-accurate prediction model, several methodologies reported in the literature were adopted. Jeddah city was chosen as the area of study to validate the proposed framework; hence, the relevant spatial and temporal data were obtained for this city at a district scale. DF has become a significant social and economic burden, particularly in Jeddah where the disease is hyper-endemic. The Saudi government has prioritised the development of an effective and accurate dengue transmission model.

Despite the efforts of previous spatial and spatiotemporal modelling studies to analyse DF patterns, a failure to establish a comprehensive framework to involve all modelling stages in addition to visualization of the improvements of models is considered a significant research gap. Collecting data of optimal quality from several sources is also challenging. Furthermore, inefficient data pre-processing will provide insufficient

modelling by showing bias or inaccurate results. Moreover, optimal pre-processing and perfect model performance require an explanation to prove the capability of the invented framework to decision-makers and stakeholders. Therefore, a detailed and accurate spatiotemporal dengue patterns process modelling methodology is proposed together with a GIS-based explanation to validate the efficiency of the proposed framework. Thus, the current research designs a novel framework that combines machine learning (ML) and GIS to develop a high-performance model for DF spatiotemporal patterns on a district scale.

This research was guided by three research questions and is intended to improve the efficiency of DF spatiotemporal prediction models. Moreover, the answers to these questions are addressed in a separate manuscript and combined in this thesis (Chapters 3, 4, and 5). Each aspect was addressed in a single research question as follows:

1. To what extent do missing values in DF data influence the performance of models developed for the spatiotemporal prediction of the disease?
2. How can cluster-based analysis be integrated to improve the performance of the spatiotemporal model?
3. Does the temporal variation affect the simulation of dengue disease risk areas?

1.6 Research gaps

Most DF investigations have focused on Latin America and Asia as these regions have the most cases (Altassan et al. 2019). However, studies are needed in other regions where DF has become a significant public health issue, but is less characteristic and may vary, such as in the Middle East (Humphrey et al. 2016). In some nations, where there are limited resources and facilities for viral testing and reporting of information, the number

of dengue cases is greatly underestimated. Thus, improving dengue risk maps for these regions is necessary (Attaway et al. 2016).

Many factors play a role in DF transfer in Saudi Arabia, including considerable numbers of migrant workers and Islamic pilgrims from countries with endemic dengue, such as the Middle East, Asia, and North Africa (Altassan et al. 2019). Hence, the impact of different factors and spatial heterogeneity should be considered when modelling dengue disease and taking preventive steps. Moreover, significant factors that affect the disease incidents should be investigated because of the lack of research on DF in Saudi Arabia. Addressing this current research gap will help to improve the health system's preparedness (Altassan et al. 2019). The DF reports, locally and globally, show that the battle against dengue is ongoing. The information available to date indicates the need for efficient vector control, careful monitoring of different variables, and epidemiological parameters to strengthen prevention (Al-Raddadi et al. 2019). While the Saudi health authority gives DF priority, epidemiological information on DF is available only for certain communities and environments (Alhaeli et al. 2016). Therefore, population-based epidemiological data are significantly absent, although it is essential to guide and track on-the-ground prevention and disease control strategies and to maximise resources in order to combat viral efficiency and vector transmission. Moreover, there is no accurate information on the risk factors for viral dissemination due to the probable progression of the epidemic in the region (Al-Raddadi et al. 2019). However, due to the dynamicity of DF spatiotemporal patterns, the right variables cannot be ascertained through only one study. The significant predictor variables are not always the same, and can change according to when and where the studies are conducted. In other words, the significant variables for a particular place and circumstances are not necessarily going to be the same variables for another place with different circumstances (Ferrell and Brinkerhoff 2018).

Hence, the prediction risk map may be more effective if environmental and socio-economic factors are considered together with physical factors which, it has been found, play a significant role in the transmission of DF. Moreover, it is difficult to understand the principal factors causing the disease and its transmissions, as the factors are interrelated and the degree of influence of each factor varies according to a region's unique climatic conditions and cultural and socio-economic context in which the virus circulates (Teurlai et al. 2015). Moreover, DF has a variety of effects on both the vectors and the disease they transmit. As a result, determining the association between DF occurrences and numerous causal variables is crucial to increase the accuracy of predicting DF (Mala and Jat 2019b). In addition to determining related data, the quality of these data is a significant issue as it affects the performance of the proposed model (Kamkhad et al. 2016). So, a more reliable early warning system for dengue prediction may be developed by understanding the crucial ranges of factors for a given study area (Mala and Jat 2019b). The accurate prediction has the potential to minimise morbidity and death associated with this condition (Buczak et al. 2014).

The literature review revealed several gaps in previous research, and these have guided the formulation of the research questions and the objectives of this study. The research gaps identified in the literature are:

1. ***Lack of sufficient studies to investigate DF spatial and spatiotemporal transmission in Saudi Arabia:*** The majority of research on DF focuses on Latin America and Asia, where the disease is most prevalent (Altassan et al. 2019; Humphrey et al. 2016). Other places, such as the Middle East where DF is a major public health issue but is less understood, and DF epidemiology may differ, have a vital need for research.

2. ***Failure to establish a comprehensive framework to improve the DF spatial and spatiotemporal modelling performance:*** Not many studies have considered developing a comprehensive framework intended to improve the accuracy of spatiotemporal models accuracy and give a better understanding of the disease patterns and associated risk factors in the presence of MD or "poor quality data". Thus, it is crucial to develop a comprehensive framework that carefully considers each analytical stage to improve the model's performance.
3. ***Failure to apply and compare several ML methods:*** Some of the previous works have applied different ML algorithms to model the disease; nevertheless, very few have integrated and compared multiple ML models to provide better accuracy and consistency.
4. ***Not much work has been done that integrates different tools and methods to generate easy-to-read maps that facilitate controlling procedures by decision-makers:*** It is difficult to predict the occurrence of infectious disease epidemics, and fully predictive technologies are still in their infancy (Ashby et al. 2017). Accurate modelling of DF might assist public health decision-makers in taking appropriate actions to reduce the disease's impact (Buczak et al. 2012). Moreover, maps that illustrate the risk areas can be easily interpreted by policymakers and health authorities, and provide clear explanations that can help to prevent and/or control the epidemic (Naish et al. 2014b).
5. ***The lack of simulation tools for proper analysis at different modelling stages:*** Although, such tools contribute to a better understanding of the spatiotemporal patterns of diseases, which help decision-makers to create different scenarios for more efficient planning and decision-making processes to control potential future threats (De Lima et al. 2016).

In a way of summary, through the study of the previous literature, it was found that most of the studies focused on determining the risk map for the study area and the factors related to the disease in that area and the correlation of these factors with the proposed model without going into detail in the precise details related to the data, its quality, and its impact on the performance of the prediction model as Figure 1.1 illustrates. Filling these knowledge gaps is useful in improving DF spatiotemporal prediction models.

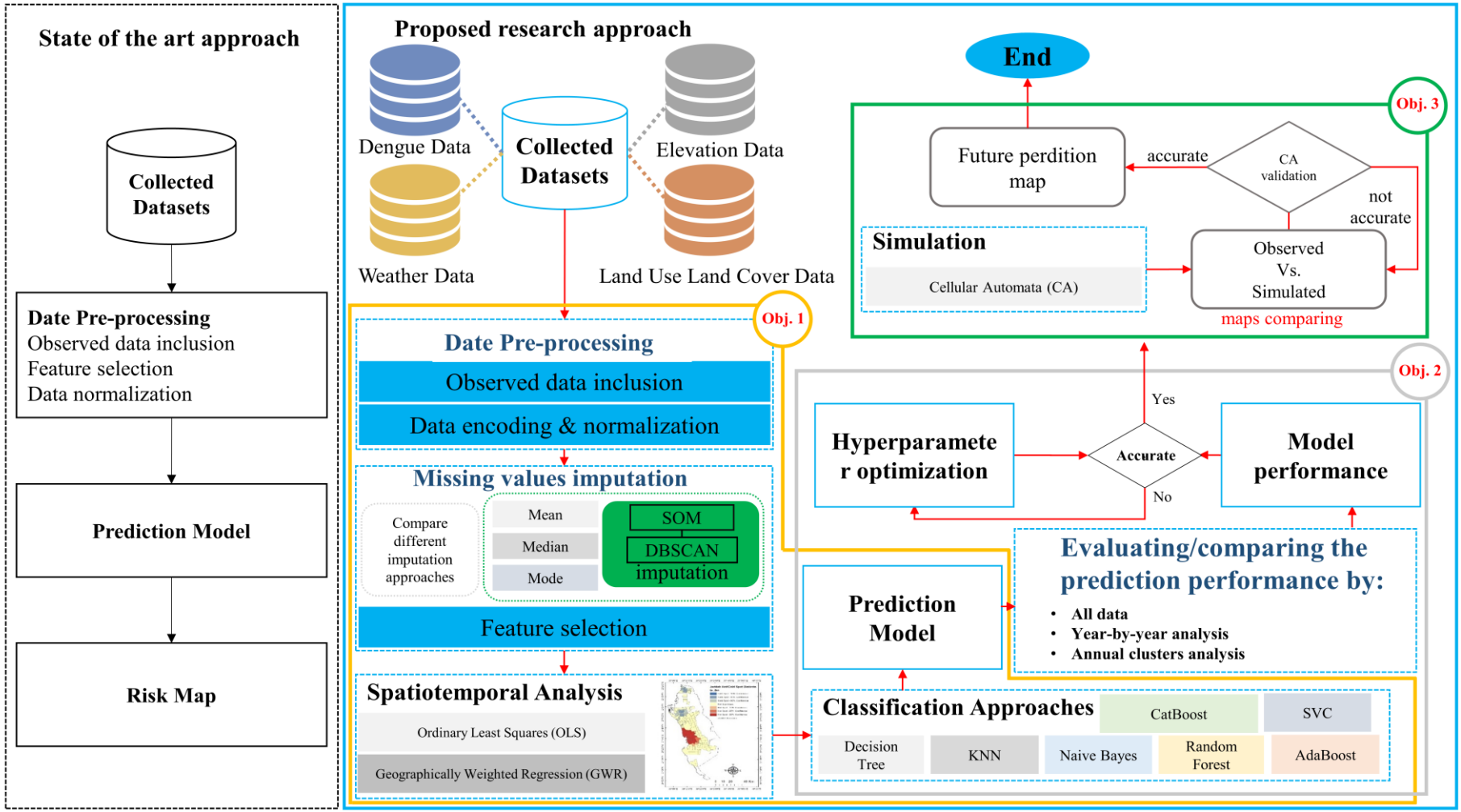


Figure 1.1. State-of-the-art research vs. proposed approach for DF spatiotemporal prediction modelling

1.7 Scope of study

An improved model with optimal performance can be developed by studying the history of disease transmission and identifying the major associated risk factors (Dhewantara et al. 2019). In addition, disease transmission is influenced by a complex combination of environmental, geographic, and socio-economic variables (Teurlai et al. 2015). This necessitates precise data on recorded case distributions, epidemiologic characteristics, and dengue transmission patterns (Valles et al. 2019). Additionally, the quality of these data determines the level of performance of the prediction model (Kamkhad et al. 2016). Therefore, the complex interplay of factors and potential danger to people and countries, contributed to the creation of a research environment to investigate, understand, and improve DF control procedures. Due to the aforementioned complexity, several technologies and algorithms, including the geographical information system (GIS), remote sensing (RS), and machine learning, have been used for spatial and spatiotemporal analysis to discover geographical relationships between disease patterns and data gathered from local governments and remote sensing satellites (Dhewantara et al. 2019; Espinosa et al. 2016). Furthermore, technologies such as maps and visualisation assist policymakers and public health authorities to communicate complicated information to the public and policymakers in an easily interpretable way (Naish et al. 2014b). Lastly, by selecting the most appropriate methods and techniques, measurements, and most informative features, it is possible to achieve better prediction performance. Regardless of the complexity of the disease and its associated risk factors, it is crucial to have precise data that allows the accurate identification of the main factors influencing the disease in order to understand its spatiotemporal patterns, develop prediction models, improve the model accuracy, and provide a clear explanation to health decision-makers. Therefore, the scope of this thesis encompasses:

- Exploring and identifying the appropriate variables associated with disease transmissions and modelling.
- Determining the shortcomings of current dengue spatiotemporal prediction models.
- Developing a model for accurate DF prediction and risk assessment.
- Integrating advanced technologies and approaches in the modelling process.
- Applying advanced ML methods and technologies to achieve better analysis and modelling.

In this thesis, the epidemiologic data used for analysis was obtained from secondary data sources. The Vector-borne and Zoonotic Diseases Administration of Jeddah in Saudi Arabia's Ministry of Health provided data for the period 2012 to 2018 for all the reported cases in Jeddah. According to previous authors (Khormi et al. 2011), the KSA vector-borne and zoonotic diseases departments responsible for health matters are keeping systematic records for all positive DF cases. The records contain numerous data including age, gender, ethnicity, neighbourhood, occupation, recording date and the disease initiation week for each case. The environmental, population, elevation and land use land cover (LULC) data were collected from several satellite images; climatic data (temperature, humidity, precipitation, and wind speed) were downloaded through Google Earth Engine (GEE) using the JavaScript export function. The population distribution was obtained from ORNL's LandScan. The Digital Elevation Model (DEM) was acquired from the US Geological Survey (USGS). LULC data was obtained from the European Space Agency (ESA) Climate Change Initiative (CCI). Then the collected DF cases were digitalised using the geographical information system (GIS) software ArcMap 10.4 to aggregate all of the parameters into one Shapefile for the analysis. Because the scope of

this thesis comprises an investigation of current modelling approaches and the development of a comprehensive analytical framework to achieve better spatiotemporal modelling accuracy, an essential step of risk assessment is to identify the principal factors associated with DF and its transmission. Therefore, for the purposes of this research, several machine techniques were adopted to achieve better results. The framework used in this thesis could indicate an evolving trend in prediction analysis and pave the way for future spatiotemporal prediction modelling.

The main focus of this thesis is to develop a comprehensive framework that could handle various analysis issues related to DF. Jeddah city in Saudi Arabia was used as the case study to validate the proposed model.

1.8 Research hypothesis

As stated previously, the main goal of the current thesis is to develop a comprehensive framework to improve DF disease spatiotemporal prediction models. Based on the literature review, three hypotheses were developed to improve DF spatial and spatiotemporal modelling. Each hypothesis was expressed and evaluated based on different analysis stages using a number of approaches. The three hypotheses:

1. The analysis of DF spatial and spatiotemporal patterns in the presence of MD affects the performance of the disease prediction model.
2. Improving the disease data analysis based on the data quality and MD using suitable imputation methods will improve the prediction performance and enhance the understanding of disease patterns where there are missing values.
3. Improving the prediction accuracy, based on the developed models from the hypothesis, will enhance the performance of simulation risk maps' accuracy

as well as highlight the main risk factors using the district boundary as a spatial scale.

1.9 Research aims and objectives

The aim of this research was to develop a comprehensive framework based on proven GIS/remote sensing technologies and ML approaches to apply the current understanding of DF epidemiology and the spatiotemporal factors contributing to this disease in endemic regions to create a simple-to-understand predictive risk map. In addition, optimization techniques were used in all stages of the analysis to obtain highly-accurate models.

It is unlikely that the transmission of vector-borne diseases will be uniform throughout large spatial areas (Tedrow 2010), since local factors determine the risk and transmission of DF (Ren et al. 2017). Thus, to meet the research aim and objectives and to understand the spatial-temporal aspects of the disease on a finer scale, the analysis of the area of interest is conducted at the district level (Jeddah City), and the period of one year as a temporal scale. This study bridges the literature gap by developing a comprehensive framework containing several complex models. Moreover, the adopted methods were integrated and improved for different purposes to achieve the thesis aims and objectives. Thus, the specific objectives of this study are:

1. to develop a data analysis model which can provide a better understanding of the incidence of DF where there is MD;
2. to improve the performance of the prediction model to achieve optimal accuracy;
3. to simulate risk maps at a district level to quantify hot spots that might be caused by imminent dengue disease.

The details of the thesis objectives are explained in the following subsections.

1.9.1 Objective 1 (Develop a data analytical model in the presence of MD)

The first objective of this thesis is to develop a data analysis model which can give a better understanding of DF spatiotemporal patterns in the presence of MD. To meet this objective, the major risk factors associated with DF transmission globally and locally (Jeddah city) were reviewed (Chapters 1 and 2). The significant factors associated with DF disease were then identified, together with their role in determining the spatiotemporal patterns of DF, and their impact on the proposed model (Chapter 3 and 4). Subsequently, a novel imputation approach was adopted to fill DF spatiotemporal missing values and improve the data quality.

1.9.2 Objective 2 (Improve MD imputation and prediction model performance)

The second objective is to improve the performance of the prediction model so as to achieve optimal accuracy. To this end, several traditional and advanced imputation approaches were adopted to fill missing values in DF spatiotemporal data and improve the data quality. A cluster-based technique was applied to determine the appropriate values that should replace the missing values. Numerous ML approaches were utilized to predict DF cases, and the performance of various models, including the one proposed in this study, was compared in terms of their prediction accuracy.

1.9.3 Objective 3 (Simulate risk areas)

After the first two objectives are met, the final objective is to simulate risk maps that offer stakeholders an easy-to-understand visualization of potential disease threats, enabling authorities to determine the districts at risk. The two parts of this objective are to (i) define and differentiate the main factors in each risk district to explain disease patterns, and (ii) validate the risk map based on the identified factors. The fulfilment of

the objectives will lead to a comprehensive framework that can accurately model the spatiotemporal patterns of DF. Moreover, the proposed model will be comprehensive and accurate, detailing all stages of the analysis.

1.10 Significance of the research

Given the capabilities of different technologies, including GIS, remote sensing, machine learning, and big data, in the public health sectors, this study aims to develop a comprehensive framework to improve the approach to DF's spatial and spatiotemporal patterns modelling. Moreover, utilizing the aforementioned techniques to output models with a high level of accuracy. Furthermore, the content of this work will connect the various modelling stages. The proposed framework can enable researchers to better understand the interactions between stages in the modelling process. Moreover, the integration of appropriate technologies and methods, taking into account all the analysis stages and choosing a model that fits the data, allows future researchers to produce models with a high level of accuracy. Hence, the proposed framework could be applied to other VBDs for better modelling and understanding of spatiotemporal patterns.

Effective and efficient surveillance systems are essential for controlling dengue virus transmission using spatial and temporal scales (Racloz et al. 2012). Moreover, combining GIS and remote sensing technologies with statistical analysis may improve the DF early warning systems (Restrepo et al. 2014). Thus, the DF surveillance and monitoring, by means of spatial and temporal scales, are essential at this stage to improve prevention and control particularly because there is no vaccine or effective treatment for this disease (Racloz et al. 2012). Moreover, visualization technology such as GIS is commonly used as an effective tool to map the DF and other VBD diseases, and improve the disease early warning systems (Restrepo et al. 2014). Additionally, in the analysis of

patterns with diverse characteristics, GIS can help researchers and health practitioners to better comprehend spatial and temporal interactions linked to disease distribution and prevention. Satellite images are also used as they are a flexible means of obtaining relevant data. Applying the aforementioned advanced technologies can ensure the coverage of the area of interest and, together, provide complete data.

This thesis addresses several aspects of DF transmission that play a significant role in modelling spatiotemporal patterns of DF disease taking Jeddah city in Saudi Arabia as a study area and improving modelling concepts in general. This study's findings will be valuable to decision-makers in Saudi Arabia, particularly the various departments of Vector-borne and Zoonotic diseases administration of Jeddah as they may further explore the spatiotemporal aspects of the disease and its associated factors. The study's outcomes may incentivise the government to support the health sector more, particularly in taking proactive preventative measures rather than using reactive techniques to control the endemic in vulnerable districts. In other words, it will assist the decision-makers in making accurate decisions and implementing the proper procedures to control and prevent disease transmission. This thesis contributes to academic scholarship and research by offering a comprehensive framework that involves several features that will improve the spatiotemporal model performance. Moreover, it will encourage researchers to investigate other means of improving a model's accuracy. Overall, this research project will make a significant practical contribution, first of all, to the decision-makers who, together with the health authorities, can establish effective policies to combat DF, and use the proposed framework to predict the potential transmission of DF in the country. Moreover, through an accurate risk map, they can allocate resources for efficient controlling of the disease spatiotemporal transmission on a district level. Secondly, an

improved prediction model can be of great use to the world's health organisations for risk mitigation as well as potential future risk reduction.

1.11 Novelty and main contribution of the research

In this thesis, methodologies for modelling and simulation analysis of DF spatiotemporal patterns based on multiple analysis stages are proposed. This study aims to address the lack of comprehensive frameworks and to improve DF spatiotemporal prediction model performance. As the main contribution, the novel combination of using simple and advanced techniques to improve the observed data quality to fit the desired disease spatiotemporal model will significantly increase the performance. Following improving the data quality and prediction performance, cellular automaton approaches and GIS are conducted to overcome the uncertainties associated with traditional methods and validate the simulated maps to the observed risk map. The current research is designed as a new innovative combination that includes all stages of modelling in detail to reach the highest accuracy prediction models capable of evaluating the risk areas and factors associated with the disease using district's boundaries as the spatial scale and a year interval as temporal scale. The city of Jeddah was chosen as the study area for multiple reasons as explained in the research motivation section. Figure 1.1 illustrates the overall research roadmap and state-of-the-art approaches for the proposed framework. Previous DF spatiotemporal studies provide an overview of the disease's significant risk factors and map the hotspot areas of the study area; they did not offer researchers better insights into the impact of each analysis stage (e.g., pre-processing and data quality) on the desired spatiotemporal prediction model performance.

1.12 Thesis organization

This thesis consists of five chapters, details of which are given below.

Chapter 1: The first chapter introduces DF from global and local perspectives, the research background, problems statement, research motivation, the research questions, research gaps, the scope of the study, the hypothesis of current work, objectives and aims, followed by the research novelty and scientific contribution and, lastly, the current section "the thesis structure in detail".

Chapter 2: Presents a literature review of relevant studies on spatial and spatiotemporal prediction models for DF. The chapter consists of three main parts. First, an examination is conducted of various common models, both simple and complex, that have been applied to the problem at hand. Second, the four most common factors (climatic, demographic and socio-economic, entomological, and environmental) that contribute to the DF disease are investigated and categorised. Lastly, all of the modelling stages from data acquisition to assessing the model performance are discussed. The chapter discuss the disease status in the study area, the current challenges in DF spatiotemporal modelling, and a general summery.

Chapter 3: Describes the methodologies employed to accomplish the research objectives, and gives details of the proposed advanced framework from which data was obtained for this research, describes the specific study area, the overall methodology, and the method applied to create highly accurate spatiotemporal prediction models for DF. This chapter demonstrates the sources of obtained data, ethical approval, the importance of considered factors, followed by descriptions of the used software and the chapter conclusion.

Chapter 4: Presents the research results and discusses them in detail. The results demonstrate the performance of the proposed approaches used to impute missing values

and improve the accuracy of DF spatiotemporal models. The chapter describes the analytical methods employed and compares several prediction models to assess their accuracy and efficiency. Future spatiotemporal disease patterns are simulated using the collected data in order to offer better controlling procedures.

Chapter 5: Concludes the thesis with an acknowledgement of the strengths and shortcomings of the research, and suggests avenues for future research in this field.

CHAPTER 2

LITERATURE REVIEW

This chapter provides an extensive review of previous spatial and spatiotemporal modelling of DF disease, assessing the different modelling stages by taking a more comprehensive approach that begins with data collection and ends with the final prediction model performance. Moreover, it discusses the study area's disease status and previously used methods. This chapter discusses various common algorithms, both traditional and advanced, that are used to model the disease. The four principal factors categories that influence dengue fever (DF) disease are explained. The data pre-processing methods and those used for imputing missing data MD to improve the model accuracy are explained. Then, the commonly-used methods applied to assess the accuracy of models are analysed. In general, this chapter offers a comprehensive view of the different modelling stages used to model DF disease spatiotemporal patterns.

2.1 Introduction

Dengue fever (DF) is the world's most prevalent arboviral infection, caused by two types of mosquitoes: *Aedes aegypti "Ae. aegypti"* and *Aedes albopictus "Ae. albopictus"* (Puggioni et al. 2020; Zheng et al. 2019). DF is widely transmitted, causing more than 100 million infections globally per year (Alhaeli et al. 2016). DF is caused by any one of the four common Dengue virus (DENV) serotypes (DENV-1, DENV-2, DENV-3, DENV-4) (Yung et al. 2015). However, dengue viruses primarily circulate between humans and vector mosquitoes, and the existence of vector viruses is a restricting transmission factor (Higa 2011; Whitehead et al. 2007). In other words, DF cannot be transmitted without the agency of an infected person or infected mosquito. Although the

transmission of this disease depends on certain factors such as particular climatic conditions, including temperature, humidity, and precipitation, the disease is found in many regions of the world (Ferrell and Brinkerhoff 2018).

Although the first large outbreaks of the disease which came to be known as DF were reported in 1779 and 1780 in Asia, Africa and North America, DF symptoms were reported much earlier, a Chinese encyclopaedia describing the symptoms of sickness and treatments, published during the Chin Dynasty (265 to 420), is the oldest known record of DF (Gubler 1998). However, in the intervening centuries since the first reported cases of this infection, the reported incidence of the disease has increased 30-fold and has been transmitted worldwide, particularly to most of the tropical and subtropical regions (Cucunawangsih and Lugito 2017). By 2015 the transmission rate had reached 13.68 per 100,000 individuals in most of Latin America (Altassan et al. 2019). The incidence of DF is now a global concern, found in over 100 countries and putting at risk more than 3.9 billion individuals living in areas with a strong likelihood of DF transmission (Domingo et al. 2010). Statistics show that DF appears to be the most rapidly transmitted, mosquito-borne viral sickness on Earth (Attaway et al. 2016; World Health Organization 2019), infecting 100-400 million people annually, in 250,000 of whom the disease advances to DHF/DSS, resulting in approximately 40,000 deaths (Bhatt et al. 2013; Khan et al. 2008; World Health Organization 2019). In recent decades, the disease has raised major concerns for governments and authorities worldwide as different serotypes have emerged with overlapping spatiotemporal factors (Domingo et al. 2010). Most DF research has focused on Latin America and Asia, which have a higher incidence of this disease (Altassan et al. 2019). Cases of DF have appeared in Africa, although precise data for this region are unavailable (Attaway et al. 2016). Furthermore, in the Middle East, where DF is a serious public health concern, the epidemiology of the disease may differ, and because

it is poorly understood, there is an urgent need for research (Altassan et al. 2019). Dengue infections are mostly due to the contact of vector mosquitoes with humans, and the vector's proximity is a significant transmission element (Higa 2011). Dengue transmission patterns have been studied closely in previous research using different epidemiological models to improve the understanding of the main factors conducive to a DF epidemic (Zhu et al. 2018). However, little is known about the geographical and spatiotemporal patterns of DF disease and their drivers (Ren et al. 2017; Zhu et al. 2018).

There is little doubt that DF has become a major public health problem worldwide, and the annually-increasing numbers of DF cases are of great concern to governments and public health authorities (Cucunawangsih and Lugito 2017). Thus, the prediction of DF transmission and the determination of risk factors have become essential as they will assist governments and health authorities take appropriate measures to control the disease, particularly since there is no vaccine or effective treatment available (Andre et al. 2008; Hsueh et al. 2012; Jain et al. 2019; Mala and Jat 2019b; Naish and Tong 2014; Ren et al. 2019; Sarma et al. 2020; Stanforth et al. 2016; Yu et al. 2011). Another challenge in the struggle to control and predict DF transmission is the undetected patterns of DF virus transmission in different regions and the fact that there are different serotypes that overlap spatially and temporally (Bouzid et al. 2014; Domingo et al. 2010). Therefore, the transmission of the virus can be controlled by integrating the various aspects of vector management (Machault et al. 2014). Thus, the initial step that must be taken in order to control and predict DF transmission in a specific region is to understand the main factors that influence the transmission of the disease (Mala and Jat 2019b). Generally, the transmission of DF across various geographical regions is due to a number of factors, including changes in regional climate (weather temperature, rainfall, humidity, global warming), population growth and travel, unplanned urbanisation, and pathogen

transmission (Bouzid et al. 2014; Fischer et al. 2013; Hsueh et al. 2012; Khormi and Kumar 2011; Koyadun et al. 2012; Mondini and Chiaravalloti-Neto 2008; Wen and Tsai 2016; Wu et al. 2009). Additionally, the combination of different factors can lengthen the DF-active season or become progressively more severe in countries where DF was not as evident previously (Huang et al. 2012). Therefore, investigation is needed to determine the various environmental, socio-economic, and biological factors that contribute to the transmission of DF. This will enable risk zones to be identified, which will help authorities to control the transmission of this disease (Hagenlocher and Castro 2015).

A review of the literature found that previous studies have different objectives and perspectives in regard to the modelling of DF. One study investigated the association between environmental factors and cases of urban dengue in terms of various scale sizes (home, district, and administrative) (Marti et al. 2020). Another paper examined the impact of vector indices on DF transmission (Bowman et al. 2014). Meghnath (Dhimal et al. 2015) studied the spatiotemporal distribution of several mosquito-transmitted diseases, including DF, based on the weather changes in Nepal. Due to the lack of an effective treatment for DF, governments and health authorities worldwide are becoming increasingly concerned with predicting DF and determining the associated risk factors for better surveillance and control (Hsueh et al. 2012; Jácome et al. 2019). From this perspective, Runge-Ranzinger et al. (2014) examined DF prediction tools and the monitoring of disease distribution by means of several surveillance systems in order to provide recommendations based on the findings for affected countries. Moreover, Sallam et al. (2017) and Phuyal et al. (2020) examined the spatiotemporal distribution models of DF causing mosquitoes *Ae. aegypti* and *Ae. albopictus* and the related socio-economic, climatic, and environmental factors that influence them. In addition to investigating the

main risk factors associated with DF, Aswi et al. (2019) analysed and compared Bayesian modelling methods previously applied for diseases.

In this chapter, the main objective is to investigate previous methods used by researchers for spatial and spatiotemporal modelling of DF transmission at different stages, from data collection and data quality assessment to the final model performance. The resulting framework may help to improve model accuracy and give decision-makers an adequate risk map in their effort to control the transmission of disease. The review also sheds light on the main variables affecting disease transmission, and examines the common methods used for dealing with MD while protecting its quality, the various clustering approaches, and the measurement of the accuracy of the proposed model.

2.2 Background

2.2.1 Main factors contributing to DF disease transmission in previous studies

Dengue viruses primarily circulate between humans, causing mosquitoes "vector", and the existence of vector viruses is a restricting transmission factor (Higa 2011; Whitehead et al. 2007). In other words, DF cannot be transmitted without the agency of an infected person or infected mosquito. Specifically, DF is transmitted by the bite of an *Aedes* mosquito infected with the dengue virus. Figure 2.1 illustrates the disease transmission from an affected person to an unaffected person by *Ae. aegypti* and *Ae. albopictus* mosquitoes.

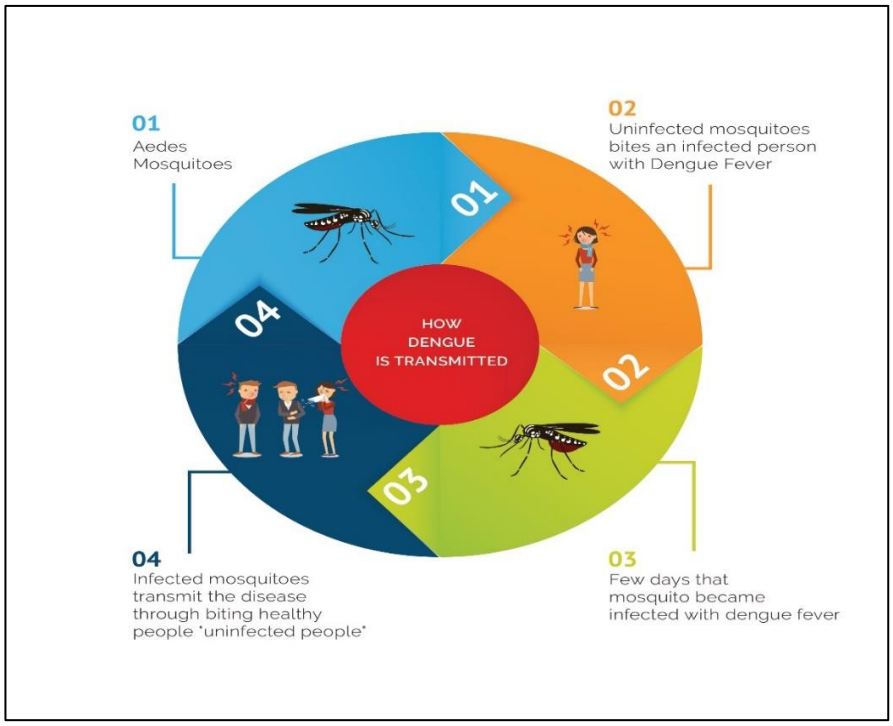


Figure 2.1 The mechanism of dengue virus transmission

Ae. aegypti is a peridomestic mosquito that lives around the abodes of humans (Moncayo et al. 2004; Powell and Tabachnick 2013), where it lays its eggs in man-made conduits and storage containers such as rubber pipes, plastic cans, bottles, drums, and concrete tanks (Arboleda et al. 2012; Mathur et al. 2018). It has been found that during the single gonotrophic cycle, *Ae. aegypti* can feed two or even three times (Mathur et al. 2018). It is mostly active and feeds throughout the day and its complete lifecycle is comprised of the egg phase to the larval stage to full adulthood (Ashby et al. 2017). Climatic "meteorological" parameters such as air temperature are of crucial importance in controlling the different life stages of *Ae. aegypti* (Acharya et al. 2016; Astuti et al. 2019). The *Aedes* mosquitoes have four lifecycle stages occurring over a period of 14 days: eggs, larva, pupa, and adult (Astuti et al. 2019). Eggs are hatched in as little as five days and form into pupae (Abou El-Saoud et al. 2018). The humidity extends the DF

vector's life and shortens the time needed for viral replication. The mosquito incubation period ranges between 3 to 14 days (Alhaeli et al. 2016).

The transmission of pathogens is susceptible to climatic conditions, particularly temperature, precipitation, and humidity (Altassan et al. 2019; Bai et al. 2013). According to a previous study (Altassan et al. 2019), the optimum weather condition for disease transmission peaks at temperatures of ≥ 30 in some places, but at an average temperature of >18 °C in other regions. Moreover, some models predict that 29°C is a suitable temperature for *Ae. aegypti* to be actively transmitting the disease, while 26°C is better for *Ae. albopictus* (Altassan et al. 2019; Mordecai et al. 2017). Also, DF mosquito activity is believed to be partly due to seasonal rain that can leave pockets of stagnant water in human habitats (Acharya et al. 2016; Astuti et al. 2019; Jácome et al. 2019). Rain or the ambient moisture itself is often sufficient to prevent the mosquito ova from being dried out (Valdez et al. 2018). Additionally, the average annual humidity was found to be the strongest predictor of DF disease (Altassan et al. 2019). Despite the temperature enabling mosquitoes to fly ranging from 15°C to 32°C, they can travel in conditions where the humidity ranges from 30% to 90% (Abou El-Saoud et al. 2018). Hence, the survival of the Aedes's mosquito depends on certain climatic conditions. In some countries, people live in crowded conditions in close proximity to massive numbers of mosquitoes. This dense population encourages the transmission of the dengue virus between humans due to the "mosquito-human transmission cycle" vector's (Gubler 2011). Moreover, the spatial transmission of DF is feasible only when either one or both of the infected organism pathogens move (Enduri and Jolad 2018).

Despite the disease circulation depicted in Figure 2.1, its transmission differs depending on spatial and temporal factors (Nguyen et al. 2020). For instance, climatic variables are known to contribute significantly to dengue transmission (Alkhaldy 2017).

Temperature can determine the length of time that mosquitoes survive, their habitats, reproduction and maturation, as well as their infective period (Wu et al. 2009). Moreover, it has been shown that the land temperature influences every stage of the mosquito's life cycle, for which the ideal land temperature is 25–30°C (Dhewantara et al. 2019). Higher temperatures can affect the DF vector's through shorten the incubation period (Carbajo et al. 2012), and may also increase human exposure to mosquitoes as people open up their houses and spend more time outdoors (Astuti et al. 2019; Hales et al. 2002; Mala and Jat 2019b; Mudele et al. 2020; Nguyen et al. 2020; Wu et al. 2009; Yu et al. 2011). Although some areas of the globe have reported higher temperatures which can encourage the survival and transmission of the vectors, at the regional level, there is limited scientific evidence that temperature fluctuations determine the extent of dengue epidemics (Wu et al. 2009). Moreover, previous studies have considered air temperature or land surface temperature, which can be obtained from remote sensing images with a smaller spatial scale that more accurately records the environmental conditions (Yue et al. 2018).

Rainfall provides the water necessary for the breeding environment of female mosquitoes and the immature larvae/pupas, leading to dengue outbreaks. Moreover, the amount of precipitation partly determines the abundance of the predominant vector, *Ae. aegypti* (Valdez et al. 2018). Rainfall can fill outdoor containers, providing a breeding site for mosquitoes (Huang et al. 2018). On the other hand, a large amount of rain can flush out overfilled containers, increasing the mortality of adult vectors and larvae (Nguyen et al. 2020; Tran et al. 2020; Yu et al. 2011). Hence, precipitation is a significant factor that affects the life cycle and population dynamics of mosquitoes and the transmission of DF (Cao et al. 2017). However, in previous global studies, there appeared to be no association between the amount of rainfall and the number of mosquitoes (Ashby et al. 2017).

Several studies have found that humidity levels positively correlate with DF transmission (Acharya et al. 2016; Astuti et al. 2019; Mala and Jat 2019b; Xu et al. 2019). High temperatures and high levels of rainfall create humidity, which provides ideal conditions for mosquito reproduction and survival, subsequently leading to the rapid transmission of DF (Hales et al. 2002). Further studies have established that humidity levels are important for mosquito abundance and reproduction: relatively low humidity and a longer dry season lead to higher egg mortality rates (Lega et al. 2017). However, where shade is being provided by canopies of fully-grown trees, there is less evaporation of the water in which eggs are hatching; therefore, the surrounding soil retains its moisture, thereby increasing the density of *Aegypti* larvae (Mudele et al. 2020).

Some global studies have considered wind speed contributing to DF transmission. However, there are inconsistent results in this regard. Previous works observed that wind speed was significantly associated with the incidents of the disease (Fairos et al. 2010; Mala and Jat 2019b), while other studies (Depradine and Lovell 2004; Rahman et al. 2018) found that the wind speed factor had a negative correlation with DF transmission. Other researchers have found that increased transmission of dengue is correlated not with climatic conditions, but with socio-economic factors (Vincenti-Gonzalez et al. 2017).

Several studies have shown that significant changes in population, transport and living conditions due to rapid urbanisation have changed or added to the factors contributing to DF transmission. Natural environmental factors such as temperature and rainfall have significant effects on the temporal and spatial trends of DF epidemics, together with socio-economic factors such as the distribution of population, people density, level of urbanisation, and road network densities (Ren et al. 2019). Also, DF transmission by infected individuals is accelerated in densely-populated regions (Ren et al. 2019). Furthermore, certain socio-economic factors (e.g., education, intelligence,

behaviour, career, jobs, means of livelihood, etc.), biological factors (e.g., age, immunity, health status) and institutional factors (e.g., health care access, quality of care, control and preventive strategy etc.), can determine people's level of susceptibility to the disease (Delmelle et al. 2016). It has been hypothesised that the increase in dengue cases is a result of rapid urbanisation and increased foreign travel in these areas (Carbajo et al. 2012). Some studies have found that the extent of urbanisation is a key factor in the transmit of DF, with proximity to the major urban centres in one province in Thailand being a key factor (Mala and Jat 2019b; Teurlai et al. 2015; Wen et al. 2015; Wu et al. 2009). Furthermore, larger cities are more vulnerable to DF transmission due to their population density and the readily-accessible containers that serve as breeding sites for *Ae. aegypti* (Carbajo et al. 2012).

2.2.2 The impact of data quality on DF modelling performance accuracy

According to Rubin (Rubin 1976), there are generally three reasons that data contains missing values: missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR) (Baraldi and Enders 2010). The three types of MD were explained by Baraldi, Kang, and Scheffer (Baraldi and Enders 2010; Kang 2013; Scheffer 2002). Although Wisniewski (Wisniewski et al. 2006) recommended seven steps to reduce the amount of MD through a well-designed and controlled study, researchers and analysts can encounter the dilemma of MD during any stage of the analysis (Pigott 2001). Moreover, both complex methods and simple traditional ones have been applied to handle the MD issue. The common traditional approach is to ignore these records and include only complete records in the analysis (Zhang 2016). Moreover, the MD can be ignored if its impact on the analysis is negligible and the amount of MD does not exceed 5% of the total data (Jakobsen et al. 2017). Single imputation using the mean

value and regression imputation is another traditional method (Scheffer 2002). Recently, a more advanced and complex imputation method known as multiple imputation (MI) has been applied to address the problem as it considers the uncertainty of the MD and provides more reliable and unbiased results (Eekhout et al. 2014).

Analysing the data distribution and MD patterns before determining their impact on the final model is often helpful in ascertaining why values are missing in the data (Richman et al. 2009). Generally, the data is noisy and incomplete; therefore, data preparation is frequently the most crucial aspect of data mining. It is estimated that data preparation accounts for up to 80% of the data engineering work undertaken to improve the model accuracy (Sessa and Syed 2016). Various methods of identification and classification of patterns, including biometric recognition, document categorisation, and medical diagnosis, have been effectively applied to resolve issues in many areas (García-Laencina et al. 2010). However, when applied to real-world categorisation tasks, pattern recognition systems often face the issue of having missing or unknown data (Brown and Kros 2003). This has given rise to an increased interest in and the application of machine learning (ML) methodologies and techniques derived from statistical learning theory (García-Laencina et al. 2010). In addition to traditional imputation methods, more sophisticated models based on ML technologies have been applied in recent years due to their flexibility and ability to capture the relationship between items in the collected data (Jerez et al. 2010). Moreover, imputation approaches inspired by ML rely on creating a prediction model to estimate the values of MD from the information contained in the dataset (Wenbai et al. 2021). Imputation methods such as multi-layer perceptron (MLP), k-nearest neighbours (KNN), self-organizing maps (SOM), and decision tree (DT) building algorithms have been widely applied in many problem areas and developing fields (Jerez et al. 2010).

Data quality is an essential issue in ML and in other related applications (Batista and Monard 2002, 2003). Therefore, an appropriate imputation approach must be adopted to ensure that the missing values do not affect the quality of desired ML models (Ramesh et al. 2021). Numerous previous studies have shown the impact of poor-quality data on the proposed model since adequate data is necessary to reflect realistic scenarios and develop effective strategies to control the disease; conversely, poor data quality leads to poor outcomes (Espinosa et al. 2016). Additionally, Luong (Nguyen et al. 2020) stressed the importance of improving data quality since the current data can model the disease in limited spatiotemporal scales. Moreover, the quality of predictors is essential in determining the variables selected for the target model (Buczak et al. 2012). Thus, MD is a significant problem adversely affecting predictions or suggestions derived from data-driven models (Batista and Monard 2002, 2003). To address this issue, previous studies have proposed several approaches for dealing with missing values in health datasets. Some studies (Fuentes-Vallejo 2017; Naish and Tong 2014; Wen et al. 2015) retained only the complete data and excluded any data with missing values from further analysis. This is an acceptable approach under some circumstances when, for example, the dataset is missing 5% or less of the total data, or the dependent parameters contain missing values that affect the identification of related independent features (Jakobsen et al. 2017). In another work (Buczak et al. 2014), authors have filled the MD values in spatial information by referring to open-source data. Moreover, the authors used linear interpolation to obtain data for missing years. Another study used the mean value to fill MD points for some features/variables, or used "zero" instead of missing values for data pertaining to rainfall, elevation, or ocean location (Buczak et al. 2012). It is evident that more emphasis should be placed on how to tackle the issue of MD when conducting research, and in the analysis of the available data to avoid the impact of such data on the

desired model and the potential bias (Kang 2013). Moreover, the accurate data will lead to accurate findings and, in turn, to more relevant recommendations (Fuentes-Vallejo 2017).

2.2.3 DF modelling, simulation and explanation

The spatial patterns of DF are complex enough to present several obstacles to dengue transmission research (De Lima et al. 2016) as it is influenced by a complex combination of environmental, geographic, and socio-economic variables (Acharya et al. 2018; Delmelle et al. 2016; McGough et al. 2021). Therefore, it is critical to understand the interplay between DF and possible risk factors (climatic, demographic and socio-economic, entomological, and environmental) which are connected to the spatiotemporal patterns of disease, and will assist authorities in taking appropriate control measures (Cao et al. 2017; Teurlai et al. 2015). Thus, an accurate and efficient prediction of the risk of dengue sickness is crucial for the monitoring, surveillance, and prevention of the disease (Faisal et al. 2010). Hence, most of the research has addressed various issues related to DF, including determinating the risk factors for dengue disease and predicting risk in DF patients (Faisal et al. 2010). Moreover, methodology such as ML models and the related learning techniques are rapidly evolving, and play an important role in many geoscience applications and remote sensings such as Self-Organizing Maps (SOM), support vector machines (SVM), random forests, and decision trees (DT) (Lary et al. 2016; Scavuzzo et al. 2018). In addition, ML approaches can involve parameters ranging from a few to several thousand for effective classification and nonlinear regression systems (Scavuzzo et al. 2018).

Modelling involves a simplification of real-world challenges and problems in order to acquire knowledge about, and a thorough understanding of, a real-world phenomenon.

The system modelling technique is often used in epidemiology models to conceptualize the epidemic process (Eosina et al. 2016). To obtain a better understanding of the dynamicity of DF transmission and simulate prediction models, most of the previous studies deployed epidemics models based on a statistical model or Ordinary Differential Equations (ODE) (Medeiros et al. 2011). However, these models are incapable of capturing the details in spatial patterns of disease transmission, or of forecasting and visualising the interactions between various features (Eosina et al. 2016; Medeiros et al. 2011). Therefore, when studying epidemics, researchers used cellular automata (CA) models to examine temporal and spatial patterns (Santos et al. 2011). Because of its computational flexibility enabling it to include precise disease features, CA has been utilised to construct a wide range of compartmental epidemic models (Pereira et al. 2021). It is well-understood that appropriate tools and methodologies for analysis are critical to achieving high forecast precision (Philemon et al. 2019). In short, computational modelling improves decision-makers understanding of the temporal and spatial patterns of epidemics, as well as guiding, testing, and modifying control tactics in simulations prior to their implementation in the real world (Lemos et al. 2017). Various computer modelling approaches have been utilized to model, simulate, and subsequently comprehend different social, environmental, biological, and other types of systems (Khalil and Wainer 2020). However, numerous studies have utilized the CA model because it is dynamic, it can be integrated with other models, and it can be modified to suit the available data (Falah et al. 2020).

Furthermore, in order to comprehend the geographical properties of a dataset, visualisation is an essential procedure for effective and easy analysis (Lukasczyk et al. 2015). Visualisation is required for detecting the epidemiology of disease patterns in a specific geographical region, forecasting disease-transmitting patterns in the following

period, and improving decision-making as a result of the model's findings, thereby helping to improve the controlling procedures (Eosina et al. 2016). Moreover, forecasting future events based on previous history is performed using various methodologies to achieve disease control planning and evaluation. Forecasts provide consumers with necessary data, preventing them from making inappropriate decisions and taking the wrong actions that may have drastic consequences in the event of a future outbreak (Philemon et al. 2019; Wu and Cowling 2018). In this regard, geographic information systems (GIS) technology and spatial statistical analysis make it possible to discover local environmental and sociological factors that indicate high-risk locations for dengue, and analyse the correlations between potential risk factors and the disease's transmission (Naish et al. 2014b; Scavuzzo et al. 2018; Wu et al. 2009), and improve the early warning systems in order to control potential future threats (Lloyd 2010; Naish et al. 2014b).

Recently, an increasing number of researchers have modelled DF transmission using the geographic information system (GIS) (Jeefoo 2012). For instance, Nakhapakorn and Tripathi (2005) used GIS to examine the relationship between certain environmental factors (rainfall, temperature, humidity, and land use/land cover types) and the DF and dengue haemorrhagic fever (DHF) transmission in Thailand. The findings indicated a relationship between land use/land cover (LULC) and disease transmission in the country. In other research, in their model, (Hsueh et al. 2012) considered the factors of population density, transportation arteries, and bodies of water. The study found that DF can be transmitted over long distances in Taiwan. However, the number of DF cases decreases when there is a greater distance between roadways and rivers. Using GIS, Khormi et al. (2011) studied DF transmission in the Jeddah province of Saudi Arabia. They determined the number of DF cases and, through applying the proposed model, they found that most of the DF transmission occurs in the middle of the city. Lastly, Abou El-Saoud et al.

(2018) used both remote sensing and GIS to model DF transmission. Additionally, to identify those areas most at risk of vector growth, these researchers investigated a number of climatic and environmental factors (rainfall, wind speed, land surface temperature (LST), vegetation cover, and temperature) as well as the population density factor. The study found that the main factors that encourage larva growth are high population density, low-lying land, high land surface temperature (LST), and proximity to construction facilities. The aforementioned studies indicate that DF transmission has been investigated from different perspectives and in the context of location environmental conditions. Therefore, prior to developing an effective prediction model, it is important to understand the significant factors that could influence the likelihood of DF transmission in the study's area of interest.

2.3 DF in Saudi Arabia

Before 1994, The Kingdom of Saudi Arabia was considered a free epidemic country of DF (Al-Raddadi et al. 2019; Alhaeli et al. 2016; Altassan et al. 2019; Shahina et al. 2009). However, the first cases (289 in total) of DF in Saudi Arabia were discovered in 1994 in Jeddah province (Alhaeli et al. 2016; Altassan et al. 2019; Badreddine et al. 2017; Fakeeh and Zaki 2001; Kholedi et al. 2012). After a decade, the disease re-emerged in the western regions of Saudi Arabia and since then, the western part of the country has been considered as an epidemic area (Alhaeli et al. 2016). Moreover, following the first reported cases in 1994, many DF cases emerged in other parts of Saudi Arabia mainly in the western province that includes Al-Madinah, Jeddah and Mecca, reaching all the way to cities such as Aseer and Jizan in the south-western region of the Kingdom (Alhaeli et al. 2016). After sample incidents in the Kingdom were examined, the three DF virus strains DENV1, DENV2, and DENV3 were identified (Alhaeli et al. 2016; Khan et al.

2008). Nowadays, DF is one of the main health problems facing the Saudi government and its health department (Altassan et al. 2019). This has motivated researchers to conduct research on the disease in different areas such as Al-Madinah (El-Badry et al. 2014), Aseer and Jizan regions (Al-Azraqi et al. 2013; Gamil et al. 2014), and Mecca (Alwafi 2013; Dieng et al. 2012; Khan et al. 2008; Shahina et al. 2009). It was found that in Mecca City, from 2006 to 2008, 159 DF cases were reported to the Ministry of Health (MOH); 67% of the cases were Saudi, two were *Al-Hajj* pilgrims, while the rest were residents of Mecca (Shahina et al. 2009). This indicated that there is a possibility of DF transmission during mass religious gatherings such as *Al-Hajj* and *Al-Umrah* (Ducheyne et al. 2018). Therefore, this massive gathering of people from many parts of the world is a severe concern for the Saudi Arabia government and the world's public health organisations (Ahmed et al. 2006). Additionally, the *Ae. aegypti* mosquito that transmits the disease is commonly found in the western and south-western parts of Saudi Arabia, especially in Jeddah city which is the main gateway for international pilgrimages to the holy city of Mecca where the *Hajj* takes place (Al-Raddadi et al. 2019; European Centre for Disease Prevention and Control 2019). Previous studies found that the Red Sea region has the highest number of reported DF cases (Altassan et al. 2019; Humphrey et al. 2016). However, most of the reported DF cases are located in Jeddah province, which is the main gateway for international pilgrimages to the holy city of Mecca (Al-Raddadi et al. 2019; European Centre for Disease Prevention and Control 2019).

Despite the presence of DF in many countries worldwide, little is known about the factors facilitating its transmission in the Middle East (Humphrey et al. 2016). Similar to other regions, in Saudi Arabia, the DF epidemic is seasonal (Altassan et al. 2019), peaking mainly between March to May "Spring season", and again in November and December but to a lesser extent (Zaki et al. 2008), indicating that climatic changes influence the

transmission of DF (Dhewantara et al. 2019). However, although most of the studies on DF have focused on the causes and factors associated with disease transmission in different regions of the world, few have investigated the Middle East region (Altassan et al. 2019). One previous local study (Aziz et al. 2014) investigated the relationship between DF and several factors including water availability, water storage, and the development of transmission networks. The impact of household access to electricity was investigated by (Al-Azraqi et al. 2013). The fast growth and development of urban centres, the proximity to the endemic DF regions, and the demographic and social changes throughout the nation have created "perfect conditions" for increasing the transmission of DF and other illnesses transmitted as a result of overcrowding, substandard housing, insufficient water, and poor waste water management systems (Altassan et al. 2019). There are also other factors that influence the transmission of DF in Saudi Arabia. For instance, in agricultural regions, crops require regular watering, which can create sustainable areas for mosquito breeding (Ashshi 2017). The topography is one of the main factors responsible for DF transmission in Jizan (Alhaeli et al. 2016). In a small village in Asser province, most of the people are farmers, live close to animals, and sleep outdoors, all of which make them targets for dengue virus infection (M Ashshi et al. 2017). In addition to the environmental factors, there are socio-economic issues such as urbanisation. The massive intra-regional commerce in the Red Sea area also contributes to the transmission of the disease, indicated by several DENV serotypes emerging in Saudi Arabia's port towns (Altassan et al. 2019). Additionally, traffic is one of the means by which the infected pathogens are transmitted from Jeddah to Al-Madinah Al-Munawwarah (El-Badry and Al-Ali 2010). Moreover, due to the limited access to water supply in most of the districts in Jeddah, residents replenish supplies by using water containers, which in turn increases the number of DF incidents in Jeddah compared to

other cities in Saudi Arabia (Khormi et al. 2011). The coexistence of several ecological variables increases the chances of detecting seropositivity (Al-Raddadi et al. 2019).

Not all factors have the same degree of influence on the transmission of DF, and vary according to place and time. For example, annually in Saudi Arabia, Muslims from all over the world make their way to this country for *Hajj* (Alhaeli et al. 2016; Ashshi 2015; El-Badry et al. 2014; El-Kafrawy et al. 2016), and the number of pilgrims from inside and outside the country is increasing yearly according to the Saudi census database (General Authority for Statistics 2018). Thus, the large number of visiting Muslims and immigrant workers from dengue-endemic nations is an essential factor to be considered in Saudi Arabia (Altassan et al. 2019). The number of pilgrims is based on the distribution of 1 per 1000 visas for each Muslim country, stipulated by the Hajj Ministry (Aleeban and Mackey 2016). Some of the Muslims from hyper-endemic infection nations are travelling to Mecca and transmitting infectious diseases to other pilgrims (Khan et al. 2008; Shahina et al. 2009). This mass gathering may increase the transmission of disease if any infection exists, and may turn quickly into a severe epidemic which will require specific infection control procedures (Memish 2002). Pilgrims who travel by air could transmit the disease to Mecca from nations around the world, and vice versa (Wilder-Smith and Gubler 2008). In addition to the risks posed by the mass gathering of pilgrims and the unavailability of vaccines for many VBDs, another consideration is that the *Hajj* season is not fixed. The date is determined by the lunar calendar, “Hijri”, which is 10-11 days shorter than the Gregorian calendar (Ahmed et al. 2006; Memish 2002). Hence, in order to have better disease control, decision-makers must take this seasonal variability into consideration and always be prepared. Additionally, the rituals of *Hajj* are conducted within a small area extending approximately 25 km from the Ka'ba to Mt Arafat, and millions of pilgrims are crowded into this relatively small area (Ahmed et al. 2006). In

short, having to live within this small area for a couple of days, the changeable weather, and the large numbers of pilgrims could facilitate the transmission of disease, particularly VBDs (Ahmed et al. 2006; Memish et al. 2003; Shafi et al. 2005). Thus, as the geographical area of the *Hajj* is small, the crowding of the *Hajj* necessitates numerous physical, environmental, and medicinal services to control potential diseases. Therefore, to control the transmission of DF in the region, additional procedures and tools are required.

The Kingdom of Saudi Arabia has taken several measures to control DF. First, the dengue surveillance program was established after the initial emergence of DF in the country in 1994, and required all public facilities to report DF cases directly to the MOH within 24 hours of case identification (Al-Raddadi et al. 2019). Second, the government ordered the regular spraying of mosquito adulticide – a particular kind of insecticide that kills the pathogen mosquitoes– and introduced a fish that feeds on mosquito larvae "*Gambusia Affinis"*, into the waterways (Altassan et al. 2019). Third, the Saudi MOH and other health organisations worldwide provide annual health guidelines for pilgrims (Al-Tayib 2019; Al Masud et al. 2018). These guidelines contain several recommendations and requirements for pilgrims before, during, and after the *Hajj*. Moreover, because of the transmission of some diseases, both communicable and non-communicable, during *Al-Hajj* seasons all pilgrims are required to have a valid international certificate of vaccination against some of the more common diseases (European Centre for Disease Prevention and Control 2019). Lastly, the Ministry of Hajj and Umrah was created by the Saudi Government to handle all *Al-Hajj* and *Al-Umrah* issues including health issues (Saad 2017). Despite previous concerns and solutions, A pressing area of concern is that just as it did in 1993, the *Hajj* in 2030 will take place in Spring, typically when DF cases are at their peak (Altassan et al. 2019). One of the Saudi

government's "Vision 2030" objectives is for the number of Muslim pilgrims to reach around five million in that same year (Aleeban and Mackey 2016). If dengue disease transmission has not been controlled by that date in Saudi Arabia and the other Muslim regions, the *Hajj* could have a much more significant effect on increasing DF transmission than it does at present (Altassan et al. 2019). Therefore, additional control techniques and accurate prediction models are required to control the potential future transmission. To this end, the MOH in Saudi Arabia is collaborating with global health organisations including the World Health Organization (WHO) and the international Centres for Disease Control and Prevention (CDC) to prevent and control the disease during the Islamic gathering in Mecca (Al Masud et al. 2018). They are seeking to provide the latest technologies and innovations to safeguard the pilgrims against disease. Although cases of DF are rare during *Al-Hajj* and *Al-Umrah*, all the solutions and procedures to control and predict the disease transmission must be considered (Shahina et al. 2009).

2.4 Previous literature on spatial and spatiotemporal modelling

A literature review was conducted to select and classify articles from the relevant literature to achieve the research objective by applying a particular search technique and specific criteria for the inclusion and exclusion of articles based on the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) Guidelines (Moher et al. 2010; Moher et al. 2009). Five different biomedical and science databases (ProQuest, ScienceDirect, IEEE Xplore, Scopus, and PubMed) were searched for articles relevant to this study. To ensure a high-quality review, the search was limited to peer-reviewed journal articles written in English and published between 2000 and 2022. The research was limited to information covering the past two decades to examine the most up-to-date modelling approaches and generalizable information that keeps the research

focused. Additionally, a manual search for relevant articles was conducted through previously published works' bibliographies. The search of the five databases yielded 7338 articles; however, several inclusion and exclusion criteria were applied to determine the articles to be reviewed. In order to meet the objectives of this review, three inclusion criteria were applied; (i) peer-reviewed articles that have appeared in journals, (ii) articles published between 2000 and 2022, and (iii) articles related to spatial and spatiotemporal modelling. Three exclusion criteria were applied; (i) non-English articles, (ii) inappropriate or unclear study design, and (iii) unclear methodology.

A total of sixty-nine (69) articles were included in the literature review as shown in Table 2.1. Of the sixty-nine articles, sixty-eight were concerned with a specific region or country; only one article took a global perspective (Hales et al. 2002). Most studies present models based on historical dengue incidents; hence, they are “secondary” datasets provided by local health sectors. The remaining studies were concerned with the vector distribution or vector breeding sites (Arboleda et al. 2012; Espinosa et al. 2016; Khormi and Kumar 2012; Khormi et al. 2011; Mudele et al. 2020; Scavuzzo et al. 2018; Wen et al. 2015; Wiese et al. 2019). Therefore, this section is divided into several subsections comprising: the previous modelling methods used, the various predictors used for the modelling, the mechanisms used for dealing with MD and, finally, the clustering of the collected data to improve model performance followed by approaches used for measuring the model accuracy.

The articles reviewed in this chapter were classified according to whether they use either ML or statistical approaches to model DF spatial and spatiotemporal transmission as illustrated in Table 2.1. Also, previous studies were analysed to determine the procedures followed by researchers when dealing with data quality, including MD imputation and preparation to fit the desired model. The clustering approaches used in

previous literature were examined. Finally, the most commonly applied accuracy algorithms were investigated to determine the best-fitting models according to the variables.

Table 2.1. List of previous related works based on spatial and spatiotemporal modelling

ID	Reference	Year	Country	Study Period	Model / Methods	Techniques Category
1	(Hales et al. 2002)	2002	Global	1975 - 1996	Logistic Regression Spatial Autocorrelation Sensitivity Analysis Global Circulation Models	Statistical
2	(Wen et al. 2006)	2006	Kaohsiung City, Taiwan	2002	Three temporal indices to evaluate the severity and magnitude of an epidemic risk for (Frequency index (α)), (Duration index (β)), and (Intensity index (γ)) The local indicator of spatial autocorrelation (LISA) Monte Carlo significance test Correlation coefficient	Statistical
3	(Rotela et al. 2007)	2007	Tartagal, Argentina	2004	Fourier Harmonic Analysis Space-Time Analysis (Spatiotemporal Clusters) Knox Test Concept Linear Pearson Correlation Coefficient	Statistical
4	(Wu et al. 2009)	2009	Taiwan	1998 - 2006	Spatial Analysis Statistical Model Kernel Estimation Principal Component Analysis (PCA) Logistic Regressions	Statistical
5	(Arboleda et al. 2012)	2011	Bello, Colombia	2002 - 2008	Ecological Niche Modelling Maxent Linear Regression The Genetic Algorithm for Rule-Set Prediction (GARP)	Statistical
6	(Jeefoo et al. 2011)	2011	Chachoengsao Province, Thailand	1999 - 2007	Temporal Analysis (Pearson Correlation Coefficient) Spatial Analysis (Risk Map, Empirical Bayes Smoothing (EBS) Standard Deviational Ellipsis (SDE) Global Moran's I Statistic (Spatial Autocorrelation) Space-Time Analysis (Space-Time Cluster Analysis) Hotspot Detection (Local indicators of spatial association (LISA) The local Getis-Ord Gi*(d) statistics) Spatial Analysis	Statistical
7	(Khormi and Kumar 2011)	2011	Jeddah, Saudi Arabia	2006 - 2010	Geographically Weighted Regression (GWR) Descriptive Analysis	Statistical

					Ordinary Least Square (OLS)	
8	(Khormi et al. 2011)	2011	Jeddah, Saudi Arabia	2006 - 2010	Getis-Ord Gi* Statistics Frequency Index	Statistical
9	(Lin and Wen 2011)	2011	Kaohsiung, Taiwan	2002	Ordinary Least Squares (OLS) Geographically Weighted Regression (GWR) Statistical Analysis Spatial Autocorrelation Coefficient (Moran's I)	Statistical
10	(Yu et al. 2011)	2011	Taiwan	2002 - 2007	Bayesian Maximum Entropy (BME) Poisson Regression Model Theoretical Covariance Models	Statistical
11	(Buczak et al. 2012)	2012	Peru	2001 - 2009	Fuzzy Association Rule Mining Logistic Regression	Machine Learning
12	(Carbajo et al. 2012)	2012	Argentina	1991 - 2011	Statistical Model (Generalized Linear Mixed Models (GLMM))	Statistical
13	(Hsueh et al. 2012)	2012	Taiwan	2003 - 2008	Hot Spot Analysis (Moran's, Local G-Statistics) Geographically Weighted Regression Model (GWR)	Statistical
14	(Khormi and Kumar 2012)	2012	Jeddah, Saudi Arabia	2006 - 2010	Hot spot / spatial clustering (Getis–Ord Gi* statistic)	Statistical
15	(Jeefoo 2012)	2012	Chachoengsao province, Thailand	2000 - 2007	Local Spatial Autocorrelation Statistics (LSAS) Kernel-Density Estimation (KDE) Methods A Risk Zone Map (Getis-Ord*i G statistic) Local Spatial Autocorrelation Statistics (LSAS)	Statistical
16	(Fan et al. 2014)	2013	Guangdong Province, China	2005 - 2011	Time-Stratified Case-Crossover Analysis Spatial Autocorrelation (Moran's I)	Statistical
17	(Buczak et al. 2014)	2014	Philippines	2000 - 2010	Fuzzy Association Rule Mining	Machine Learning
18	(Machault et al. 2014)	2014	Martinique	2009 - 2011	Decision Tree Logistic Regression	Statistical
19	(Naish et al. 2014b)	2014	Northern Queensland, Australia	1992 - 1993	Descriptive statistics (Statistical analyses) Spatial analysis Spatial Autocorrelations (global Moran's I test) Semi-variogram models Interpolations of SIR values (kriging)) Temporal analysis (chi-square analyses)	Statistical
20	(Naish and Tong 2014)	2014	Queensland, Australia	1993 - 2012	Descriptive Spatial and Temporal Analyses Hot Spots Spatial Autocorrelation (Local Indicators of Spatial Association (Anselin's Local Moran I test, LISA) Spatial Clustering Spatial and Space-Time Scan Statistical Analyses	Statistical
21	(Restrepo et al. 2014)	2014	Colombia	2007 - 2010	Bayesian Spatiotemporal Conditional Autoregressive Model, Poisson Regression Models Spearman Correlation Analyses Spatial Autocorrelation,	Statistical

					Conditional Autoregressive (CAR)	
22	(Viennet et al. 2014)	2014	Queensland, Australia	1995 - 2011	Spatial Analysis Time-Series Analysis Cross-Correlation Analysis	Statistical
23	(Ortiz et al. 2015)	2015	Cuba	1981 - 2013	Spatial Modelling Time-Series Analysis Spatial Autocorrelation (Moran's I, Local Indicators of Spatial Association LISA) Exploratory Spatial Data Analysis Spatial Autoregressive Specification (SAR)	Statistical
24	(Teurlai et al. 2015)	2015	New Caledonia	1995 - 2012	Multivariable Model, Spatial Autocorrelation (SAC) Principal Component Analysis (PCA) Support Vector Machines (SVM)	Machine Learning
25	(Wen et al. 2015)	2015	Taiwan	2009 - 2010	Negative Binominal Regression Statistical Analyses Kernel Density Mapping Regression Coefficient	Statistical
26	(Acharya et al. 2016)	2016	Nepal	2010 - 2014	Choropleth mapping technique Spatial autocorrelation (empirical Bayes approach) k-nearest neighbourhood Poisson-based model Monte Carlo simulation	Statistical
27	(Delmelle et al. 2016)	2016	Cali, Colombia	2010	Kernel Density Estimation (KDE) Spatial Autocorrelation (Moran's I) Geographically Weighted Regression (GWR) Ordinary Least Squares (OLS)	Statistical
28	(Espinosa et al. 2016)	2016	Tartagal Salta Province, Argentina	2009 - 2014	Hot spots (heatmap tool) Annual density breeding site maps (Kern density algorithm) Cluster analysis (SaTScan) 999 Monte Carlo replications Land cover classification(k-means) classifiers Ecological niche model	Statistical
29	(Manica et al. 2016)	2016	Rome, Italy	2012	Statistical Analysis Generalized Linear Mixed Models (GLMMs) Generalized Additive Mixed Models (GAMMs)	Statistical
30	(Mutheneni et al. 2018)	2016	Andhra Pradesh, India	2011 - 2013	Spatial Statistical Analysis Hotspot (Getis-Ord Gi *) Self-Organizing Map (SOM)	Machine Learning
31	(Stanforth et al. 2016)	2016	Río Magdalena, Colombia	2012 - 2014	Bayesian Estimation Model Principal Component Analysis (PCA)	Machine Learning
32	(Tian et al. 2016)	2016	Guangzhou, China	1978 - 2014	Cross-Validation Enhance Spatial and Temporal Adaptive Reflectance Fusion Model (ESTARFM) Iterative Self-Organizing Data Analysis Techniques Algorithm (ISODATA) Phylogenetic Analysis Bayesian Skyride Analysis Metropolis-Hastings Markov Chain Monte Carlo Algorithm	Statistical

33	(Zhu et al. 2016)	2016	Guangzhou, China	2014	Ross-Macdonald theory Radiation Model Markov chain Monte Carlo (MCMC) Method Mathematical Model	Machine Learning
34	(Ashby et al. 2017)	2017	Magdalena River, Colombia	2012 - 2014	Boosted Regression Tree (BRT)	Machine Learning
35	(Cao et al. 2017)	2017	Guangzhou, China	2014	Spatial Autocorrelation (Moran's I)	Statistical
36	(Fuentes-Vallejo 2017)	2017	Girardot, Colombia	2012 - 2015	Space (Getis-Ord Index) Space-Time Clusters, Spatial Autocorrelation (Moran's) Kulldorff's Scan Statistics Monte Carlo Simulations	Statistical
37	(Hafeez et al. 2017)	2017	Lahore, Pakistan	2011	Spatial Autocorrelation (Local Index of Spatial Autocorrelation (LISA))	Statistical
38	(Ren et al. 2017)	2017	Guangzhou and Foshan (GF) cities, China	2014	Geographically Weighted Regression (GWR) Spatial Autocorrelation Analysis Spatial Modelling	Statistical
39	(Vincenti-Gonzalez et al. 2017)	2017	Maracay, Venezuela	2010 - 2012	Spatial Analysis Hot Spot Detection (Risk Map) Monte Carlo Randomization procedure Point Pattern Analysis (PPA) Logistic Regression Multivariate Analysis Local Spatial Statistics	Statistical
40	(Zellweger et al. 2017)	2017	Noumea, New Caledonia	2008 - 2009 and 2012 - 2013	Spatial Analysis Statistical Modelling Multivariable Generalized Linear Model Sensitivity Analysis Spatial Autocorrelation (Moran's I Statistic, a Local Indicator of Spatial Association (LISA)) Multivariable Negative Binomial Regression Models	Statistical
41	(Zhu et al. 2018)	2017	Guangdong, China	2014	Wavelet Approach (Morlet Wavelet) Linear Regression Meta-Population Model Mathematical Models	Statistical
42	(Acharya et al. 2018)	2018	Jhapa district, Nepal	2011 - 2016	Spatial Autocorrelation (Global and Local Moran's I) Ordinary Least Square (OLS) Geographically Weighted Regression (GWR)Semiparametric Geographically Weighted Regression (s-GWR) Pearson Correlation	Statistical
43	(Huang et al. 2018)	2018	Taiwan	2014 - 2015	Statistical Analysis Spatial Autocorrelation Analysis (Moran's I) Spearman's Rank Correlation Coefficient Generalized Linear Mixed Models (GLMMs) Descriptive Statistics Sensitivity Test and Stratified Analysis	Statistical
44	(Jácome et al. 2019)	2018	Guayaquil, Ecuador	2000 - 2017	Spatial Autocorrelation (Moran's I)	Statistical

					Descriptive Statistics, (Maxent) Model Spatial Distribution Model Partial Least Squares Regression (PLS-R)	
45	(Mala and Jat 2019b)	2018	Delhi, India	2006 - 2015	Histogram Analysis Poisson Regression	Statistical
46	(Ong et al. 2018)	2018	Singapore	2006 - 2016	Statistical Analysis Random Forest	Machine Learning
47	(Scavuzzo et al. 2018)	2018	Tartagal city, Argentina	2012 - 2016	Correlation Matrix Support Vector Machine Artificial Neural Networks Linear Regressions (Simple Linear Models, Ridge Linear Models) K-Nearest Neighbours Support Vector Regression (SVR) Multilayer Perceptron (MLP) k-Nearest Neighbour Regression (KNNR) Decision Trees Regression (DTR)	Machine Learning
48	(Yañez-Arenas et al. 2018)	2018	Mexico	2002 - 2016	Regional and Global Niche Models Principal Components Analysis (PCA) Statistical Validation Maxent	Machine Learning
49	(Yue et al. 2018)	2018	Guangzhou, China	2014	Spatial Pattern Analysis, Spatial Autocorrelation (Global Moran's I) Hot Spot Analysis (Getis-Ord Gi*) Spatial Statistical Models Spearman Rank Correlation Ordinary Least Squares (OLS)	Statistical
50	(Akter et al. 2019)	2019	Queensland, Australia	2010 - 2015	Linear Regression Models Time Series Seasonal Decomposition Analysis Space-Time Cluster Analysis (SaTScan) Descriptive Analysis	Statistical
51	(Dhewantara et al. 2019)	2019	Bali, Indonesia	2012 - 2017	Seasonal Trend Decomposition Analysis with Loess (STL) Smoothing Bayesian Spatial and Temporal Conditional Autoregressive (CAR) Modelling (Bayesian Spatial Model) Descriptive Analysis Spatial Autocorrelation (Moran's I Analysis, Anselin's LISA analysis) Spearman Correlation Analysis	Statistical
52	(Astuti et al. 2019)	2019	Cirebon, Indonesia	2011 - 2017	Seasonal Decomposition Analysis with Loess (STL) Cross-Correlation Analysis Generalized Linear Model (GLM) Partial Autocorrelation Function (PACF) Empirical Bayes (EB) Moran's I and Local Indicator of Spatial Association (LISA) Analyses	Statistical

					Descriptive Statistics Spearman's Correlation	
53	(Jain et al. 2019)	2019	Bangkok, Thailand	2008 - 2015	Statistical Analysis, Generalized Additive Models (GAM) Poisson Regression Quasi-Poisson Regression	Machine Learning
54	(Liu et al. 2019)	2019	Guangzhou & Guangdong, China	2011 - 2015	Generalized Additive Mixed Model (GAMM) autocorrelation (ACF) partial autocorrelation functions (PACF) Spearman's rank correlation coefficients	Statistical
55	(Raju et al. 2019)	2019	Kerala, India	2011 - 2016 and 2018	Correlation Analysis Cross-Validation Linear Regression Support Vector Regression K-means Clustering Linear Kernel Logistic Regression Naïve Bayes	Machine Learning
56	(Ren et al. 2019)	2019	Guangzhou, China	2012 - 2017	Geographically Weighted Regression (GWR) Spatial Autocorrelation Analyses (Moran's) Ordinary Least Square (OLS)	Statistical
57	(Valles et al. 2019)	2019	Quezon City, Philippines	2010 - 2015	Correlation Analysis Self-Organizing Map (SOM) Ordinary Least Squares (OLS) Geographically Weighted Regression (GWR) Random Forest Regression Spatial Autocorrelation k-means-clustered	Machine Learning
58	(Whiteman et al. 2019)	2019	Panama	2005 - 2017	Discrete Poisson Space-Time Modelling STSS Poisson Space-Time Modelling	Statistical
59	(Wiese et al. 2019)	2019	Pennsylvania, USA	2001 - 2015	Species Distribution Modeling (SDM) Parameter-Elevation Regressions on Independent Slopes (PRISM) Analytical climate model Pearson correlation analysis	Machine Learning
60	(Xu et al. 2019)	2019	Thailand	1999 - 2014	Distributed Lag Non-Linear Model Spatiotemporal Pattern Analysis Spatial Cluster Analysis Poisson Regression Model Generalized Linear Model Multivariate Meta-Analysis	Statistical
61	(Zhou et al. 2019)	2019	Guangzhou, China	2014	Geographically Weighted Poisson Regression (GWPR) Analysis of Variance (ANOVA) Exploratory Analysis Spatial Autocorrelation Analysis	Statistical
62	(Akter et al. 2021)	2020	Queensland, Australia	2010 - 2015	Multivariate Poisson Regression Models Bayesian Markov Chain Monte Carlo Spearman Correlation Analyses	Statistical
63	(Mudele et al. 2020)	2020	Espírito Santo State, Brazil	2017-2019	Linear correlation coefficient (R) Spatial autocorrelation Random forests (RF) regression Maximum Entropy methods Support Vector Regression (SVR)	Machine Learning

					Decision Trees Regression (DTR) k-nearest Neighbor Regression (KNN) Artificial Neural Networks (ANN) Multilayer Perceptron (MLP) Statistical Regression Models Linear Regression Model (LM) Generalized Linear Model (GLM)	
64	(Puggioni et al. 2020)	2020	Puerto Rico	1990-2014	linear Regression Moran's I Bayesian	Statistical
65	(Nguyen et al. 2020)	2020	Central Vietnam	2005 - 2018	Statistical Analysis Generalized Additive Model (GAM)	Statistical
66	(Tran et al. 2020)	2020	Reunion Island	2012 - 2013	Process-Based Approach k-fold Cross Validation Technique Support Vector Regression (SVR) Spearman's Correlation Coefficient	Machine Learning
67	(Francisco et al. 2021)	2021	Manila, Philippines	2012-2014	Random Forest Pearson Correlation Analysis	Machine Learning
68	(Sriklin et al. 2021)	2021	Pattani, Yala, and Narathiwat provinces, Thailand	2015-2019	Spearman Rank Correlation Poisson Regression Analysis Multivariable Poisson Regression	Statistical
69	(Prasetyowati et al. 2021)	2021	Jakarta, Indonesia	2007 - 2018	Spearman's Rank Correlation Moran's I Local Indicator for Spatial Association Analysis Classification and Regression Tree (CART)	Machine Learning

Due to the complexity of DF, and because there is no effective medicine or vaccine that can prevent its transmission, appropriate modelling approaches are a means of acquiring an understanding of the disease's spatiotemporal transmission patterns and controlling the incidence of DF (Hsueh et al. 2012; Jain et al. 2019; Mala and Jat 2019b; Naish and Tong 2014; Sarma et al. 2020; Stanforth et al. 2016; Yu et al. 2011). Several studies have been conducted globally to model DF transmission and to understand the main factors affecting this transmission. Figure 2.2 shows previous research that modelled dengue transmission in particular research areas in attempts to understand the spatiotemporal transmission patterns of the disease and to identify the main factors causing the disease in the study areas in order to control potential future threats.

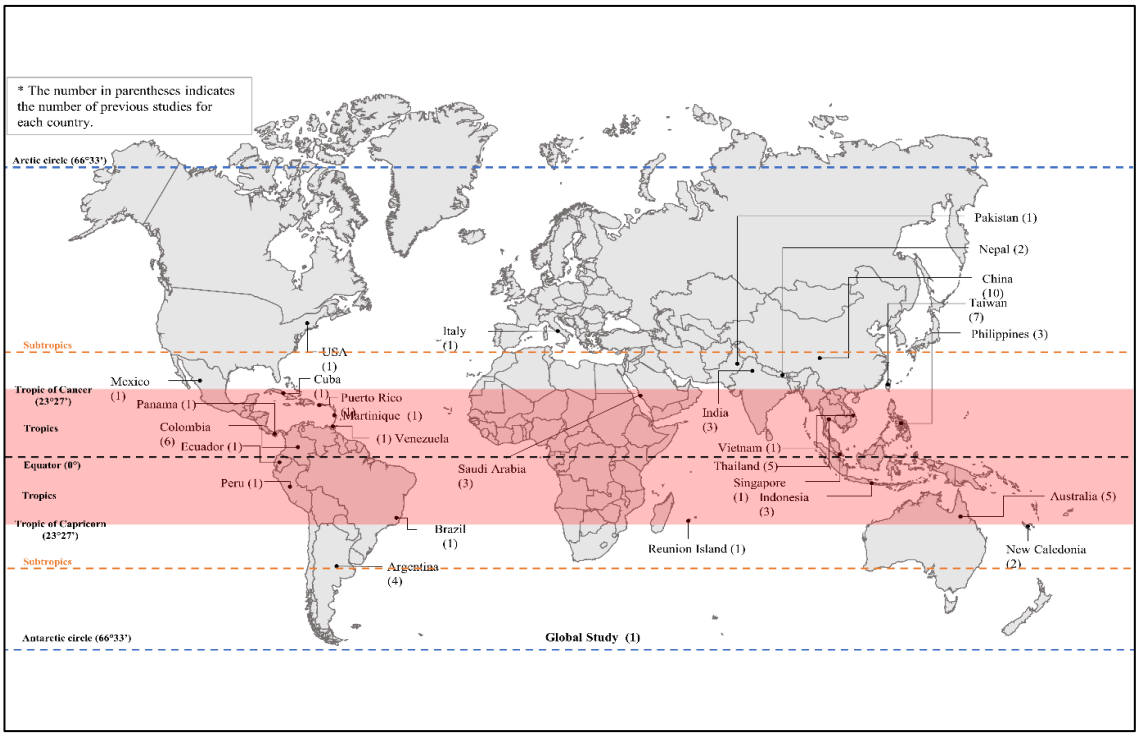


Figure 2.2. Previous dengue fever transmission studies worldwide

2.4.1 Techniques used for spatiotemporal modelling

Various studies have modelled DF transmissions in spatial and spatiotemporal scales using two main techniques, namely ML and statistical methods. Despite the fuzzy line between statistics and machine learning, some methods can be used in either domain for prediction and inference to achieve the intended goal, whereas the two have a different focus (Xu and Jackson 2019). A traditional statistical method relies on assumptions about how the data are generated, while ML involves the creation and development of algorithms whose performance improves with experience (Ij 2018). Moreover, ML is more flexible than the basic statistical model and can capture interactions within the data, which in turn improves the accuracy of the prediction model (Cauthen et al. 2016). Recently, several ML approaches were considered for modelling and predicting future disease transmission. These included Self-Organizing Map (SOM), Random Forest (Ong et al. 2018), Support Vector Machine (SVM) and k-Nearest Neighbour (k-NN)

Regression (Scavuzzo et al. 2018), and the Fuzzy Association Rule Mining technique was used for prediction modelling (Buczak et al. 2014; Buczak et al. 2012). On the other hand, statistical analysis methods are commonly used for modelling health problems and do not require the large datasets required by the ML method. Moreover, statistical analysis requires certain information about the predictors/variables for the modelling process (Henley et al. 2020). Therefore, in this thesis, the articles were grouped into two categories based on the type of technique used by the researcher(s) in previous works. As illustrated in Figure 2.3, the ML approach is the least popular method used for this type of modelling, while statistical techniques are the most popular. This indicates that most of the previous studies tend to focus on the final model and not address the approach taken to deal with the collected data and the methods applied to clean the data and impute missing records before finalising the model. The following subsections describe these study categories in detail and the corresponding methods used.

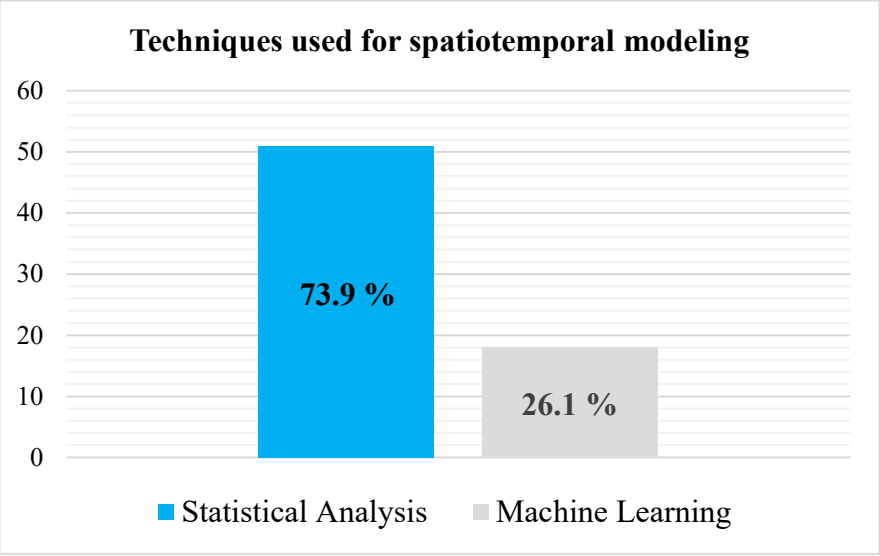


Figure 2.3. Common analysis techniques in literature

It is essential to improve dengue risk maps, particularly for developing nations with limited resources for viral testing and few information reporting and surveillance facilities, especially where the number of dengue cases is greatly underestimated (Attaway et al. 2016). Therefore, firstly, it is important to determine the principal factors influencing the transmission of DF in the area of interest. Transmission is influenced by climatic, entomological, socio-economic, and environmental factors (Akter et al. 2017). To begin with, the climatic factors in different regions play a significant role in DF transmission. These factors include sea surface temperature (SST), land temperature, precipitation, humidity, rainfall, wind speed, vegetation, and global warming (Dasgupta et al. 2019; Freeze et al. 2018). Secondly, socio-economic factors significantly impact DF transmission; these include human behaviour, human movement and travel, income, and urbanisation (Laureano-Rosario et al. 2018; Nakvisut and Phienthrakul 2018). Any additional factors are grouped under other categories based on the information from the reviewed articles.

Since missing values are a recurrent issue with the management of databases for their respective time series and, therefore, affect the data quality, several techniques have been created to solve this problem and obtain the missing values (Batista and Monard 2003; Silva and Zárate 2014). The issue of MD is a common problem in various scientific fields (Kang 2013). Therefore, the different methods used to address the MD problem are discussed in this chapter. The review found several studies that reported information missing from their collected data, although they did not discuss the methods they used to deal with this problem. A few studies did explain how the problem was addressed, although there was a variety of solutions including ignoring the MD and not including it in the analysis, replacing it with default values, using a custom approach to fill in the data

such as using the average for specific datasets, or applying the linear regression method as discussed later.

Clustering is an essential technique in data mining methods and is used to extract information from big data. Moreover, it requires categorising and grouping the data into multiple subsets or "clusters" based on their similarity (Saxena et al. 2017). However, to extract useful information from spatiotemporal data, both the spatial and temporal neighbours of all the objects must be taken into account (Rao et al. 2012). Thus, spatiotemporal data is more complicated than other data types (non-spatial or non-temporal) (Birant and Kut 2007). The clustering technique is applied by many to determine the risk zones for study areas, mainly using the Getis-Ord Gi* and Local Indicators of Spatial Association (LISA) approaches.

Assessing the model accuracy is an essential step in determining the best-fitting model for the data (Tedeschi 2006). Additionally, modelling in general and ML techniques in particular have been shown to be a powerful means of solving complex problems with a simple explanation of the algorithms for inexpert users (Biamonte et al. 2017). Therefore, decision-makers must not be tricked by persuasive, yet inaccurate clarifications (Papenmeier et al. 2019). Moreover, they should be aware of the extent to which the developed model will work for them. Therefore, model evaluation is necessary before it is presented to decision-makers (Ribeiro et al. 2016). The most popular mechanisms used to measure the model accuracy from analysing the reviewed papers were identified. The most commonly-used approach is the Akaike Information Criterion (AIC), followed by Receiver Operating Characteristics (ROC) and Area Under Curve (AUC), respectively.

By way of summary, numerous spatiotemporal modelling approaches have been applied to model and determine the risk factors associated with DF in specific areas of

interest. In this chapter, the previous modelling methods used to model the transmission of DF were reviewed. Based on previous modelling studies, two categories of modelling approaches were identified: statistical and machine learning. The risk factors associated with the disease and used as predictors to model the disease were then determined. The approaches that previous researchers used to impute missing values so as not to affect the quality of the collected data were then examined. Lastly, the literature on common clustering approaches and purposes were investigated.

2.4.1.1 Spatiotemporal modelling based on statistical analysis

The review found that statistical analysis is the technique most-commonly used to model DF. The majority of reviewed literature used the statistical modelling approach, and showed that various approaches have been used for spatial autocorrelation analysis of the relationship between predictors and DF in specific areas (Acharya et al. 2018; Astuti et al. 2019). This method is an effective means of measuring, spatially the correlation between selected factors. Moreover, Moran's *I* (Jeefoo et al. 2011; Puggioni et al. 2020), is a common and widely-used indicator of spatial autocorrelation (Acharya et al. 2018). Moreover, Moran's *I* value ranges from -1 indicating a perfect negative spatial autocorrelation and +1 which reflects a perfect positive spatial autocorrelation. Furthermore, values near 0 indicate that the data is randomly distributed (Astuti et al. 2019). Spearman's rank order is used in different studies (Prasetyowati et al. 2021; Sriklin et al. 2021) to determine the strength between two parameters when the data is not normally distributed (Huang et al. 2018). Pearson correlation (Francisco et al. 2021; Wiese et al. 2019) is another common method used to measure the spatial correlation for the data variables. Lastly, the partial autocorrelation function (PACF) is used to

investigate the relationship in a time series analysis with particular lagged values (Astuti et al. 2019; Liu et al. 2019).

Moreover, the regression method is another common statistical approach to modelling DF. In particular, Geographically Weighted Regression (GWR) was the most common approach used to model DF spatially and develop a risk map (Acharya et al. 2018; Delmelle et al. 2016; Hsueh et al. 2012; Lin and Wen 2011; Ren et al. 2019; Ren et al. 2017; Valles et al. 2019), followed by the Ordinary Least Square method (OLS) (Acharya et al. 2018; Khormi and Kumar 2011; Lin and Wen 2011; Ren et al. 2019; Valles et al. 2019; Yue et al. 2018). Generally, GWR models estimate a parameter of interest under more localized conditions by taking into account the location of the observation compared to the OLS regression model (Nazeer and Bilal 2018). Four studies applied Poisson regression (Acharya et al. 2016; Mala and Jat 2019b; Restrepo et al. 2014; Xu et al. 2019) as well as logistic regression (Hales et al. 2002; Machault et al. 2014; Vincenti-Gonzalez et al. 2017; Wu et al. 2009). Most often, rate analysis is conducted using Poisson regression, while proportion analysis is performed using logistic regression (Imrey 2000). Bayesian modelling is also a statistical approach to modelling DF as it expresses all uncertainty in the model based on statistical "probability" (Akter et al. 2021; Yu et al. 2011). Moreover, it is becoming increasingly popular for spatial analysis of diseases to employ Bayesian spatial models, as they reduce estimated variance, especially for regions with small populations. In addition, the model incorporates higher levels of variance, and a more comprehensive assessment of prediction uncertainty can be made by using maximum likelihood (Akter et al. 2021).

Lastly, Maxent software can be applied to model the distribution of dengue vectors using either statistical or ML techniques as a non-random relationship between two data sets is detected by this algorithm (Espinosa et al. 2016). Moreover, it is a popular

algorithm that requires the presence of species data to model the disease distribution (Wiese et al. 2019). Four of the reviewed papers applied this software as a statistical method (Arboleda et al. 2012; Espinosa et al. 2016; Jácome et al. 2019; Yañez-Arenas et al. 2018), while (Mudele et al. 2020; Wiese et al. 2019) studies used it as a ML approach due to its capability to fit highly complex responses (Wiese et al. 2019). Two studies used this approach to model the spatiotemporal location of *Ae. aegypti* larva and to model the key factors that facilitate the transmission of the disease and the movement of the vector (Arboleda et al. 2012; Espinosa et al. 2016). *Ae. albopictus* is another main vector that causes DF, and was studied by Yañez-Arenas et al. (2018) to map the environments conducive to mosquitoes or to predict future sites based on the region's characteristics.

Clustering is a technique whereby data with similar features are collected in a unified set, as well as determine risk areas (hot/cold) where usually hot "red" colour represents the high-risk areas while cold "blue" colour represents the safe areas. The spatiotemporal cluster analysis method was found to be an effective method for analysing and illustrating the relationship between the time and the location of the disease transmission (Lin and Wen 2011; Whiteman et al. 2019). Monte Carlo simulation is an assessment technique used to assess the significance level of the predictors used in the model (Akter et al. 2021; Fuentes-Vallejo 2017), while Monte Carlo randomization is used as an evaluation method to compare the outcomes of the statistical analysis (Vincenti-Gonzalez et al. 2017; Wen et al. 2006). This method is used to evaluate the statistical significance of spatial clusters due to its ability to determine the likelihood that clusters would arise by chance in a given population (Fuentes-Vallejo 2017).

2.4.1.2 Spatiotemporal modelling based on ML techniques

Several papers used ML and data mining techniques in the reviewed literature to model the DF disease. ML algorithms are a class of Artificial Intelligence (AI) science that enables a machine to learn and predict future incidents from several datasets (Alzubi et al. 2018). Moreover, ML uses data mining and other learning algorithms to create models of events according to certain information in order to determine future outcomes (Wang and Wang 2015). Moreover, data mining is the process of discovering hidden knowledge within massive databases (Ahmed and Hannan 2012).

In recent years, these fields have attracted increasing attention and are now applied in a broad spectrum of therapeutic areas such as disease classification, discovery of anti-viral drugs, disease symptoms diagnostics, other medical studies, and different fields. The traditional sequential mining of patterns is used to extract the most frequently-related objects over time (Fathima and Manimeglai 2015; Jainul Fathima et al. 2019; Mekha et al. 2016; Srivastava et al. 2020). However, ML and data mining methods have seldom been implemented to determine the spatiotemporal factors related to the transmission of DF (Flamand et al. 2014). Despite the advantages offered by machine learning, statistical analysis was utilised more often in the reviewed papers.

The review revealed that numerous ML methods are used for different purposes. For instance, Self-Organizing Map (SOM) (Mutheneni et al. 2018; Valles et al. 2019) and Principal Component Analysis (PCA) (Teurlai et al. 2015; Yañez-Arenas et al. 2018) are common machine-learning approaches used to reduce the volume of data in order to simplify the explanation of specific predictors, which would otherwise be complex. SOM aids in cluster formation, but it cannot actually identify or extract clusters. However, a cluster can be extracted from a SOM using a clustering method such as K-means or DBSCAN (Shukla et al. 2018). K-means is a common clustering method used in ML due

to the simplicity of its concepts. The K-means method was applied in a study after the data dimensions were reduced by means of SOM (Valles et al. 2019) to extract the clusters, and with Naïve Bayes in another study (Raju et al. 2019). The SOM method is easier to implement than the GIS method (Valles et al. 2019), and can easily be integrated with ArcMap software. The reduction of the volume of data facilitates the clustering of risk factors and helps to locate the hot spots in areas of interest (Mutheneni et al. 2018).

Several approaches were used for the prediction model, including Support Vector Machines (SVM) (Teurlai et al. 2015). Support Vector Classifier (SVC) was adopted in one comparison study and achieved almost 93% accuracy, which is better than other methods compared in the same study, i.e., linear regression, logistic regression, and naïve Bayes (Raju et al. 2019). Another common SVM method is Support Vector Regression (SVR), which was used in two studies (Mudele et al. 2020; Scavuzzo et al. 2018). However, this method did not perform as well as other methods in these two studies since Random Forest performed better than SVR in the (Mudele et al. 2020) study, while nearest neighbour regression (KNNR) performed better than SVR in (Scavuzzo et al. 2018). Random forest is a popular ensemble learning approach commonly used for regression and classification and as a predictive model for the risk area. According to (Ong et al. 2018) Random Forest produced high accuracy risk maps, with more than 80% of observed risk ranks falling within the 80% prediction range. The boosted regression tree (BRT) is similar to the random forest method, based on a decision tree and was applied in one reviewed paper that found this approach valuable to DF ecological niche modelling (Ashby et al. 2017). Moreover, the results illustrated that population density is a significant variable in modelling the disease. The classification and regression tree (CART) was used by (Prasetyowati et al. 2021) to determine how the value of a particular variable could be anticipated in a particular cluster. The results showed that the age and

occupation variables are important variables in the disease clusters. A specific data mining method known as Fuzzy Association Rule Mining, a set of data mining methods that automatically extract data, were applied in two studies to classify the DF (Buczak et al. 2014; Buczak et al. 2012). The main advantage of this method is its ability to be easily understood by humans using common linguistic terms. The two studies achieved a specificity accuracy of 0.982 and 0.974 respectively. The Maxent machine-learning algorithm was used in one study to assess the relative importance of environmental and socio-economic "neighbourhood" factors in predicting the presence of *Ae. albopictus* using three models based on the factors importance, accuracy and the mosquitoes predicted spatial distribution (Wiese et al. 2019). *Ae. albopictus* DF mosquitoes were predicted in urban centres to varying degrees by all three models. Moreover, compared to the model with environmental variables only (73.5%) and the model with neighbourhood factors only (72.1%) separately, the combined model had the highest accuracy (74.7%). Thus, according to the study finding, Maxent can be used to incorporate neighbourhood factors associated with the presence of vectors, complementing and improving species distribution models.

Lastly, a combination of statistical and/or ML approaches was applied in a few studies to develop and compare DF models. One study (Mudele et al. 2020) used a combination of ML algorithms to model the vector oviposition stage to determine the time when these vectors were able to transmit the disease, as well as compare ML performance to statistical models such as linear regression and generalized linear models. The ML methods include Support Vector Regression (SVR), Random Forest, k-Nearest Neighbours (KNN), and Decision Trees Regression (DTR). The results show that ML approaches outperform statistical models and the random forest method provides the best performance. In another comparative study (Scavuzzo et al. 2018), four ML approaches

namely support vector machines, artificial neural networks, K-nearest neighbours, and a decision tree regression model were compared against two linear approaches. A comparison of ML approaches to linear approaches clearly shows that ML approaches perform more effectively than linear approaches, particularly nearest neighbour regressions (KNNRs). Using three ML algorithms, Linear Regression, SVR, and Kernel Ridge, a comparison study was performed on the meteorological and geospatial analysis of the collected data, and among all algorithms, SVR performed the best (Raju et al. 2019). Figure 2.4 shows all of the mechanisms used to model DF in the reviewed literature.

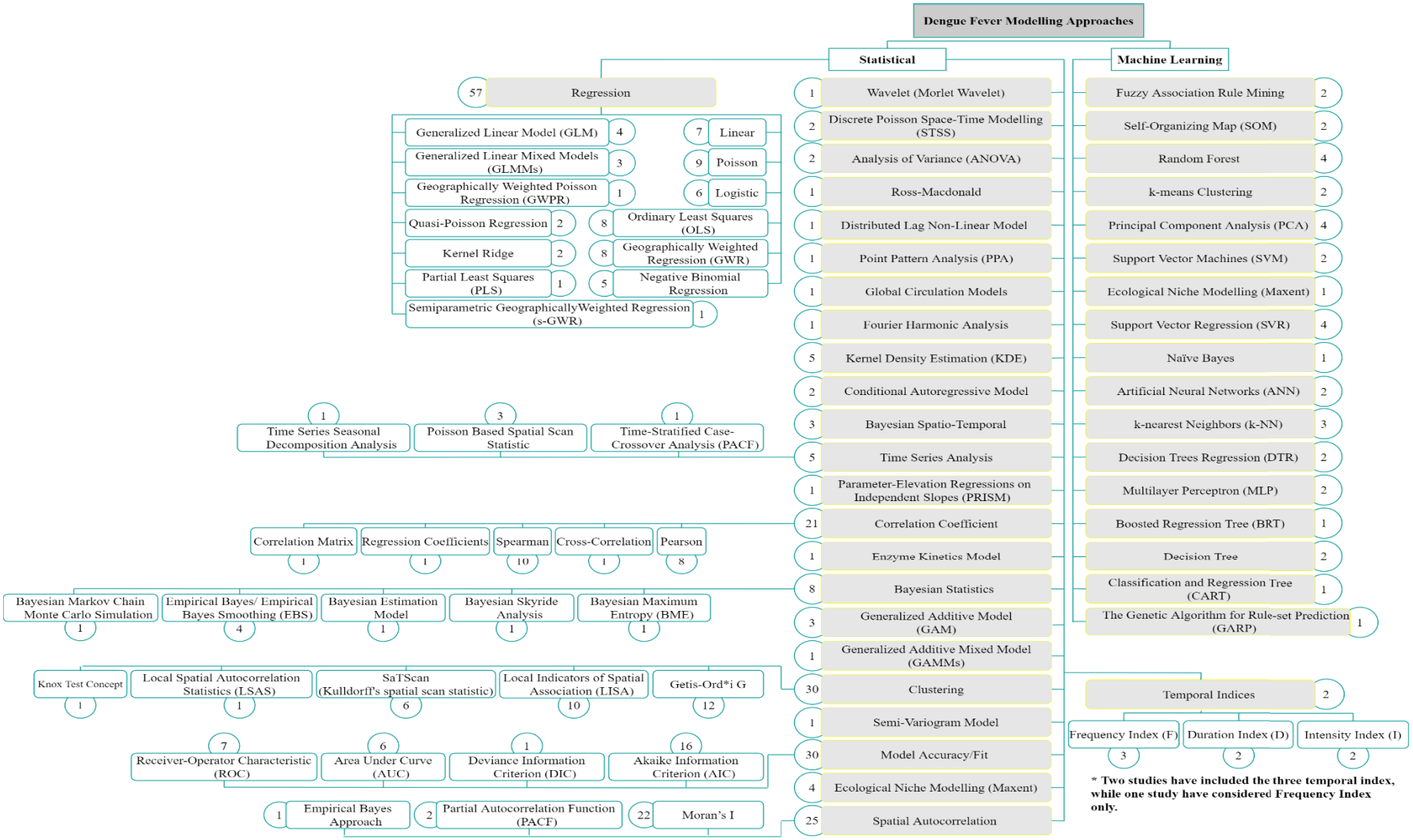


Figure 2.4. Dengue fever modelling approaches reported in the literature

2.4.2 Main risk factors “predictors” used for previous DF transmission modelling

The main risk factors associated with DF transmissions are based on the study region and differ from one geographical area to another (Jácome et al. 2019). As a means of dealing with the global risk of dengue virus disease, predictive models will enable health organisations to control the disease and eliminate future threats (Teurlai et al. 2015). Recently, many researchers have applied statistical, ML algorithms and data mining techniques to create significant predictive models as explained previously. Prior to modelling DF transmission effectively, it is necessary to determine the disease's spatial and temporal characteristics to define the main predictors (Naish and Tong 2014). Thus, the common predictors identified in the literature were grouped into four main categories: climatic, demographic and socio-economic, entomological, and environmental (Table 2.2). Each of these is discussed in the following sub-sections.

Table 2.2. Variables considered in reviewed papers

Factors	Articles ID																																																																				
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56	57	58	59	60	61	62	63	64	65	66	67	68	69	Total
Climatic																																																																						
Temperature	✓			✓	✓	✓				✓		✓			✓	✓	✓	✓		✓	✓	✓	✓	✓				✓		✓	✓	✓	✓		✓	✓		✓			✓	✓	✓	✓	✓			✓		✓	✓	✓	✓	✓	✓				✓	✓		✓		✓		✓	✓	✓	✓	43
Rainfall	✓			✓	✓	✓				✓	✓				✓	✓	✓					✓		✓				✓	✓	✓			✓				✓						✓		✓					✓	✓		✓		✓		✓					✓			✓	✓		✓		27
Humidity						✓									✓	✓		✓				✓	✓														✓	✓				✓	✓		✓		✓				✓	✓		✓	✓					✓								✓	✓	19
Precipitation												✓									✓		✓					✓			✓			✓	✓	✓		✓					✓	✓			✓	✓				✓		✓					✓				✓	✓			✓		✓	20
Oceanic Niño Index															✓			✓																																																				2
Southern Oscillation Index (SOI)										✓	✓					✓	✓																																																				✓	5
Indian Ocean Dipole (IOD)																																																																				✓	1
Sea Surface Temperature Anomaly											✓						✓																																																					2
Land Surface Temperature (LST)																													✓					✓													✓		✓								✓						✓		✓					7
Atmospheric Pressure																✓							✓																																													✓		3
Typhoon Status and Wind Unisys Weather																	✓																																																					1
Wind Speed												✓																		✓														✓	✓																									4
Dew Point												✓																																																										1
Sea Level Pressure												✓																																																										1
The Surface Reflectance Daily																																		✓																																				1
Mean Diurnal Range																																																✓																						1
Isothermality																																																✓																						1
Average Sunshine Hours																																													✓																									1
Water Vapor Pressure	✓																						✓																					✓																										3
Solar Radiation																							✓																					✓																										2
Dissolved Oxygen Density in Air																							✓																																															1
Insolation																							✓																																															1
Flood																																																																			✓			1
Demographics & Socio-Economic																																																																						
Population Data (age, sex, nationality)	✓	✓					✓	✓			✓				✓	✓	✓		✓			✓				✓				✓	✓			✓			✓		✓	✓					✓												✓	✓				✓			✓					22
Population Density				✓		✓	✓				✓		✓		✓									✓			✓		✓		✓		✓	✓	✓		✓			✓	✓		✓		✓	✓			✓		✓	✓				✓	✓		✓		✓	✓								27
Population Size													✓						✓	✓												✓	✓					✓			✓		✓																					✓						9
Population Distribution				✓								✓			✓																					✓																					✓													5
Crowded Places																									✓																																													1
Age Structure				✓																	✓						✓																																											3
Ethnic Groups																												✓																			✓																							2
Political Stability											✓						✓																																																					2
Gross Domestic Product (GDP)																											✓		✓						✓			✓		✓	✓		✓						✓							✓														9
Access to Electric Lighting											✓																																																✓											2
Housing																								✓															✓	✓		✓																	✓		✓	✓								7
Occupation																								✓															✓	✓																													✓	4
Land Urbanization Level (LUL)					✓		✓																				✓		✓						✓			✓																		✓														7
Running Water											✓																																																											1
Hygienic Services											✓																																																											1
Household Density				✓																				✓															✓	✓																			✓								✓		✓	7
Human Mobility																								✓									✓																											✓									3
Road Density																																						✓																													✓		✓	3
Blood Samples																✓																																																						1
Transportation																																	✓																							✓														2
Garbage Collection																																																				|✓|																	1
Sanitation																	✓																																										✓											2
Education																											✓													✓																														2
Internet Home Access																																								✓																														1
Water Storage																																							✓																														✓	2
Travel / Immigrant Population																												✓													✓	✓					✓			✓																				5
Connectivity																																										✓																												1
Poverty																																																											✓											1
Entomological																																																																						
Breteau Index (BI)				✓	✓	✓				✓																		✓				✓	✓																																✓					8
Mosquitos Positive Houses																																																										✓												1
Breeding Percentage (BP)																																														✓																								1
House Index (HI)																												✓																																						✓					2
Container Index (CI)																		✓																																																✓					2
Immature Stages						✓												✓																																																	✓				3
Mosquito Density Index (DI)																																																																	|✓|				1
House Infestation Index (HIF)																																																																	|✓|				1
Breeding Site					✓																				✓			✓																																										3
Egg Density Index																																															✓																							1
Infestation Index																							✓																																															1
Mosquito Samples / Mosquitoes Data (Vector Population Data)								✓						✓																																													✓				✓							4
Ovitrap Indices (Ois)																																																																			✓			1
Environmental																																																																						
Land Use and Land Cover (LULC)																		✓										✓	✓	✓	✓	✓											✓						✓							✓			✓			✓								11
Enhanced Vegetation Index (EVI)											✓						✓																	✓																									✓								✓			5
Normalized Difference Vegetation Index (NDVI)					✓						✓						✓	✓																	✓			✓		✓		✓	✓		✓	✓	✓	✓				✓				✓					✓		✓							17
Normalized Difference Water Index (NDWI)																		✓																													✓		✓										✓				✓							5
Adapted NDWI Mac Feeters Index (ANDWI)																		✓																																																				1
Normalized Difference Built-Up Index (NDBI)																																										✓																												1
Water Body																																											✓													✓					✓									3
Elevation					✓						✓						✓				✓										✓																				✓	✓							✓											8
Proximity to Epidemic Canals																											✓																		✓																									2
Slope																																																											✓											1
Flow Accumulation																																																											✓											1

* This listing order of categories is based on considered variables of each group from the most considered to the least considered variables and it indicates the frequency of used variables in the reviewed paper and does not necessarily imply the order of importance regarding significance as determinants.

2.4.2.1 Climatic factors

The climatic factor is an appropriate and standard variable to begin examining the relationship of any disease transmission resulting from climate change (Naish et al. 2014a). Climatic predictors, more so than any other, are the most commonly-used variables. Additionally, numerous climatic predictors are considered as principal factors that affect disease transmission. The temperature predictor in particular is widely used to model the transmission of DF. Here, the temperature is split into two different predictors based on the method used to measure it. Data for the temperature predictor is usually obtained from the local meteorology department in the area of interest, while the Land Surface Temperature (LST) is the temperature data obtained from satellite images. Moreover, LST has measurement criteria that are different from those of metrological stations; therefore, it is considered as a separate variable. Temperature is the most commonly-used predictor in the reviewed papers, followed by rainfall, then both humidity and precipitation. Other factors that are considered include the oceanic Niño index and the sea level pressure, depending on the area being studied (Jeefoo 2012; Machault et al. 2014). Lastly, since climate changes are affected by time and location, this category is considered essential when modelling the DF. All the climatic predictors and their associated articles are presented in Table 2.2.

2.4.2.2 Demographic and Socio-economic factors

The demographic and socio-economic factors are those related to populations and communities in general. DF can be transmitted more rapidly in crowded areas and older districts with poor infrastructure; population density is a common predictor in this category (Akter et al. 2021; Zhou et al. 2019). Population density is the main demographic predictor, and was a factor that featured in most of the reviewed articles (Table 2.2).

Usually, the disease transmission depends on the number of people within the area of interest, since it was found that crowded places increase the probability of coming into contact with dengue vectors (Wen et al. 2015). Moreover, this study examined the correlation between the *Aedes* mosquitoes and crowded gathering places such as schools, malls, and cinemas. The findings showed that people who gathered in crowded areas have more chance of being bitten by mosquitoes. Therefore, in addition to the population density predictors, population size and distribution have an impact on the transmission of the DF disease (Fuentes-Vallejo 2017; Valles et al. 2019). The population data for this variable is usually obtained from records of patients who had DF symptoms and tested positively or negatively (Khormi and Kumar 2011). This variable is considered to be a predictor since it shows the geographic distribution of patients with the disease.

DF is more common in poor neighbourhoods, and usually appears close to old infrastructures and areas with limited access to water (Ortiz et al. 2015). Therefore, all of these predictors are based on social and economic factors. Gross Domestic Product (GDP) measures individual income and is one of the main predictors of disease transmission as it can determine people's quality of life in a particular area or district (Ren et al. 2019; Yue et al. 2018). Additionally, apart from GDP, there are other predictors such as individuals' occupations (Prasetyowati et al. 2021; Zellweger et al. 2017), or the number of people living in poverty (Wiese et al. 2019). Since the disease is more common in poorer districts, numerous variables, such as housing type, were used as predictors to determine the quality of the environment, and model the disease (Akter et al. 2021; Zhou et al. 2019).

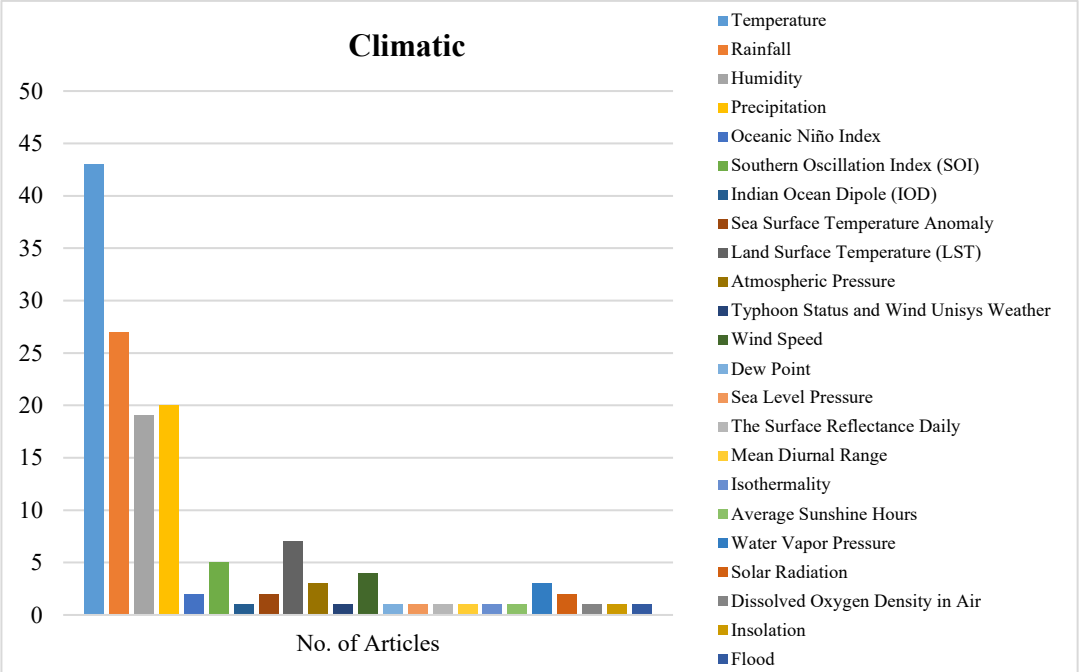
2.4.2.3 Entomological factors

Entomological predictors are related to the different life-cycle stages of *Ae. aegypti* or *Ae. albopictus*. Several indicators have been used to detect the presence of mosquitoes as well as their population density. The Breteau Index (BI) is a commonly used indicator to determine the number of mosquitoes in a specific location. Moreover, the BI index measures mosquito density in several containers (Wu et al. 2009). Since the transmission of DF is based on the presence of mosquitoes that cause the disease, other indexes are used to calculate the vector's presence throughout the various stages of its life-cycle. Additional predictors were examined in the modelling process such as House Index (HI) (Espinosa et al. 2016; Nguyen et al. 2020), Container Index (CI) (Machault et al. 2014; Nguyen et al. 2020), Egg Density Index (Scavuzzo et al. 2018), and Infestation Index (Ortiz et al. 2015). Breeding sites are related positively to human behaviour as was assessed in three studies (Arboleda et al. 2012; Espinosa et al. 2016; Wen et al. 2015). *Aedes* mosquitoes are found in man-made constructions such as tire places, outdoor water storages, and water containers (Arboleda et al. 2012; Espinosa et al. 2016; Wen et al. 2015) all of which can be breeding sites carrying the possibility of disease transmission.

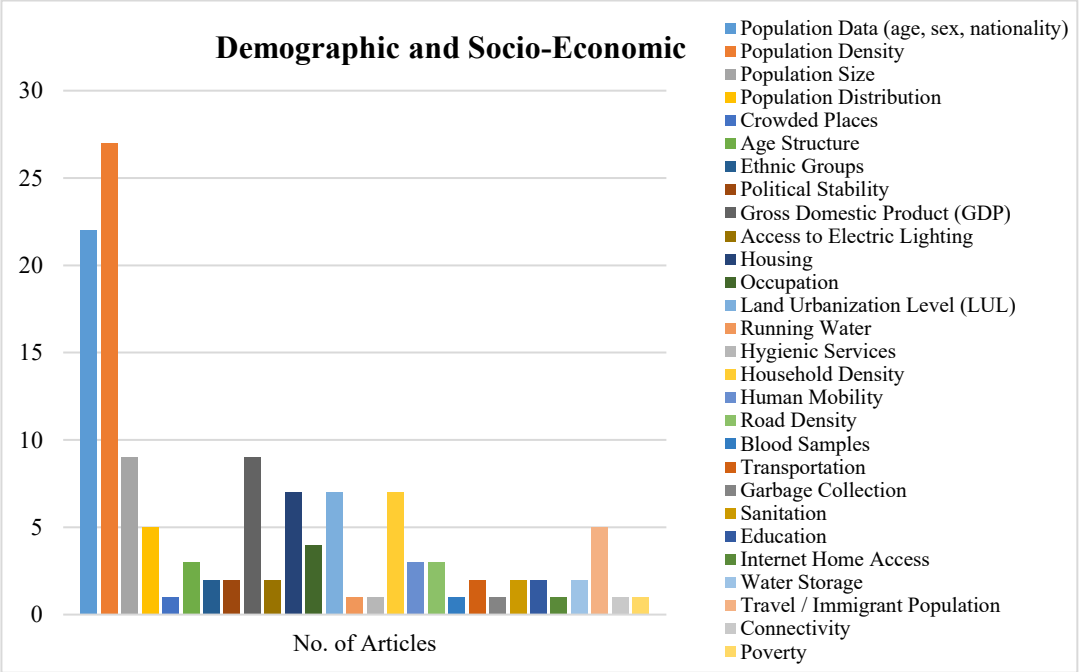
2.4.2.4 Environmental factors

In addition to the fact that human behaviour and activities may produce sites for mosquito breeding, other environmental factors provide an appropriate environment for the continuation of mosquito life and the transmission of disease (Mala and Jat 2019b). Environmental factors are other common predictors of disease transmission. Generally, land use and land cover are common predictors used to model the transmission of DF; data is usually obtained from satellite images and contain numerous predictors (Akter et al. 2021; Ren et al. 2019; Wiese et al. 2019). Most of the environmental data was obtained

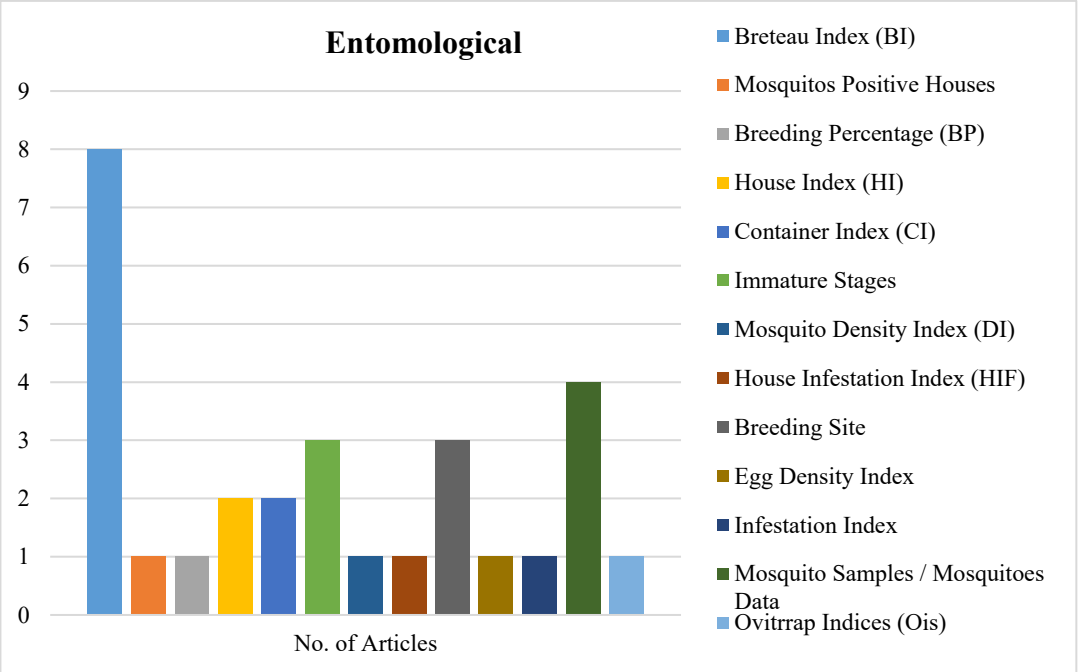
through satellite imagery, or was calculated using Geographical Information Systems (GIS). The appropriate environment for mosquitoes is often linked to three main factors, namely the vegetation, water, and land elevation as illustrated in Table 2.2. Therefore, these three predictors appear most often in the dengue models after the Normalized Difference Vegetation Index (NDVI) predictor (Mudele et al. 2020; Zhou et al. 2019). The Vegetation Index was used in almost one-third of the reviewed papers as (NDVI) or as Enhanced Vegetation Index (EVI) (Buczak et al. 2014; Buczak et al. 2012). Satellite images captured water areas and referred to as Normalized Difference Water Index (NDWI) (Machault et al. 2014; Scavuzzo et al. 2018; Yue et al. 2018). Moreover, the water predictor was extracted from the OpenStreetMap tool (Zhou et al. 2019), obtained from the local mapping department "Taiwan National Land Surveying and Mapping Center (NLSC)" (Huang et al. 2018). The elevation is another common predictor to model the risk factors related to DF disease, usually obtained from satellite images (Arboleda et al. 2012; Astuti et al. 2019; Stanforth et al. 2016; Wiese et al. 2019), in particular from the Shuttle Radar Topography Mission (SRTM). Moreover, the elevation data for one study was obtained from three local weather stations in Bali (Dhewantara et al. 2019), online using WorldClim in one study (Restrepo et al. 2014), and from a local centre website “NOAA National Geophysical Data Center” (Buczak et al. 2014; Buczak et al. 2012). Figure 2.5 shows all the factors used in the reviewed papers and presents a bar chart showing the number of times each factor was considered.



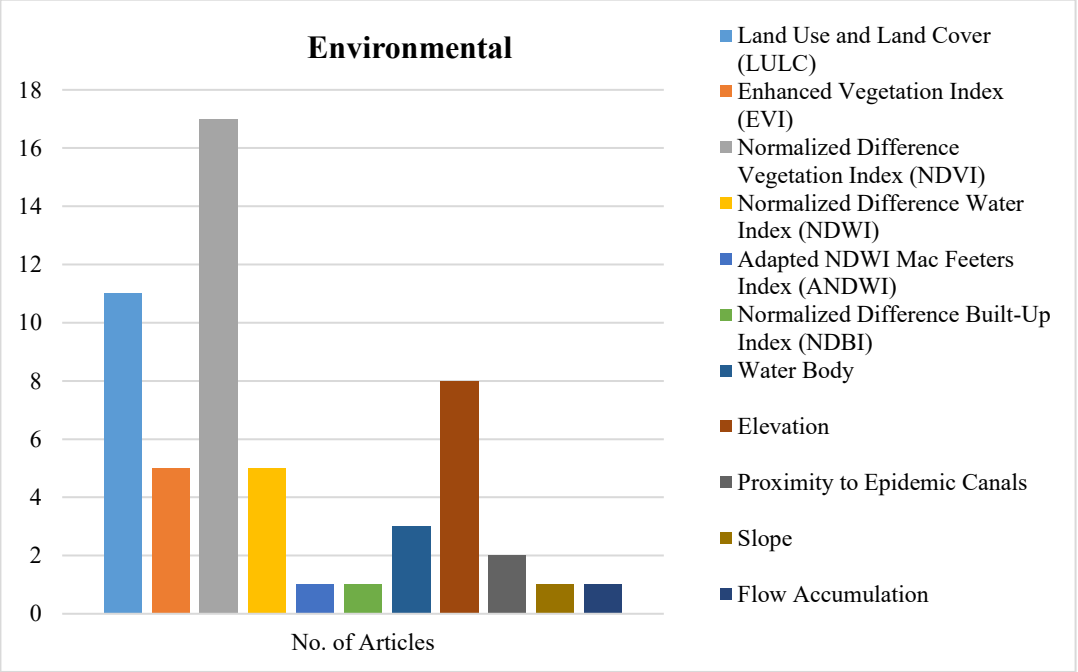
(a) Climatic factors



(b) Demographic and socio-economic factors



(c) Entomological factors



(d) Environmental factors

Figure 2.5. Principal factors for dengue fever transmission

2.4.3 Approaches used for missing data imputation

Missing or incorrectly entered data in the database is a dilemma faced by the researcher during the data analysis stage and is a general problem in most studies. Ideal modelling is essential to solving the issue (Khan 2021). Still, the problem of MD and methods of dealing with it may reduce the effectiveness and quality of the study (Kang 2013). This issue reduces the accuracy of the data and leads to inaccurate or false results, thereby compromising the validity of any conclusions. By investigating this issue in previous literature, it was found that only a few researchers touched on this problem in their studies and described how this dilemma was resolved. For example, (Zhu et al. 2016) applied stepwise linear regression to estimate MD using the Breteau Index (BI) predictor. Moreover, Hales et al. (2002) and Zhou et al. (2019) mentioned the MD issue but did not suggest any mechanisms for dealing with it. There are several approaches for handling MD, such as ignoring the predictors MD from the analysis, as seen in two reviewed papers (Mutheneni et al. 2018; Naish and Tong 2014), calculating the average for a set of data that can be used to address the problem of MD and replacing the data with results that are closer to the truth, or adopting an advanced method such as the Bayesian simulation method which is based on an iterative process to supply MD (Soley-Bori 2013). However, replacing missing values with default values may adversely affect the final readings (Kang 2013). One study (Buczak et al. 2012) adopted a particular mechanism for imputing the missing values; for instance, temperature attributes were captured from satellite images and set as “Missing” when the temperature value was missing from an entire grid cell. Other studies that were missing values for rainfall parameters, assigned totals of zero for the missing records (Mutheneni et al. 2018; Naish and Tong 2014; Zhu et al. 2016). When geographic information was missing, the researcher filled the data gap with corresponding open-source data, and replaced the data measurement of daily

incidents data with weekly incidents data to reduce the impact of MD (Zhu et al. 2016). Eliminating MD and maintaining continuous values for the study period is another procedure followed to ensure a better analysis of the data (Teurlai et al. 2015; Wen et al. 2015). The Generalized Linear Mixed Model (GLMM) was used by (Carbajo et al. 2012) since this methodology can handle the errors in collected data. Replacing the missing values by measuring the average of related data is another mechanism used to fill in the MD (Xu et al. 2019). The mathematical approach of "fifth-order spline interpolation" was applied to fill the MD in the collected datasets in one study (Mudele et al. 2020).

2.4.4 Approaches used for data clustering

Data clustering is a basic approach used to identify the hot/cold spots of DF incidents and group similar incidents in a unified group in GIS. Therefore, to obtain a prior visualization of the hot/cold areas, different methods are used to perform spatial autocorrelation to assess the relationship between each predictor and DF. The reviewed articles revealed the use of numerous methods for clustering the collected data. However, global and local Moran's I (Local Indicators of Spatial Association (LISA)) are the most common approaches used to assess spatial clusters (Dhewantara et al. 2019; Jeefoo et al. 2011; Wen et al. 2006). Getis-Ord General G (Getis-Ord*i G) is another common statistical approach used to identify the data clusters by testing the null hypothesis and visualizing the hot/cold spot areas (Hsueh et al. 2012; Jeefoo 2012; Jeefoo et al. 2011; Khormi and Kumar 2012; Khormi et al. 2011; Mutheneni et al. 2018; Vincenti-Gonzalez et al. 2017; Yue et al. 2018). Different tools such as the Optimized Hotspot Analysis tool in ArcMap (Valles et al. 2019), or the Kern density algorithm in QGI software can be applied (Espinosa et al. 2016). The spatiotemporal clustering can be performed using SaTScan software (Acharya et al. 2016; Akter et al. 2019; Espinosa et al. 2016; Jeefoo et

al. 2011; Naish and Tong 2014; Xu et al. 2019). The Self-Organizing Map (SOM) was applied to enhance clustering analysis (Naish and Tong 2014; Valles et al. 2019), as well as the k-means approach (Espinosa et al. 2016; Raju et al. 2019). Local Spatial Autocorrelation Statistics (LSAS) (Jeefoo 2012) and the Knox method were applied by (Rotela et al. 2007) to detect the spatial-temporal clusters.

2.4.5 Model accuracy methods

After the data analysis, the model's accuracy must be determined as a final step. Moreover, in order to ensure the accuracy of a model's prediction, its appropriateness must be assessed. This is crucial to establish confidence when choosing other models (Tedeschi 2006). However, researchers may find it difficult to determine the model fit depending on the complexity of the proposed model, incorrect model specifications, and the size of the data (Henley et al. 2020). This assessment stage can be conducted using either statistical or ML approaches. Seven methods were used to determine the final model accuracy as shown in Table 2.3. However, in regard to statistical models, the Akaike Information Criterion (AIC) is the mechanism most commonly used in previous studies to measure the best-fitting model, having been applied sixteen times (Machault et al. 2014). The Deviance Information Criterion (DIC) is similar to AIC and is used to compare Bayesian models (Meyer 2014), and was applied by (Dhewantara et al. 2019). R-squared ($R^{2)}$ is another statistical method that is widely used in DF modelling to show model accuracy in terms of the dependent and independent variables (Althouse et al. 2011). Receiver-Operator Characteristic (ROC) / Area Under Curve (AUC) are the second most commonly-used methods in the reviewed studies, being applied seven and six times respectively. These methods were applied in statistical (Machault et al. 2014; Wu et al. 2009) and ML approaches (Jácome et al. 2019; Stanforth et al. 2016). However,

there are several other measures that can be used to find the best-fitting model, such as Standard Root Mean Squared Error (SRMSE), and Root Mean Squared Error (RMSE) (Althouse et al. 2011). Ultimately, the choice of the mechanism depends on the type of collected data and the model being applied.

Table 2.3. Methods used to determine model accuracy

Method	Times Used
Akaike Information Criterion (AIC)	**16**
Area Under Curve (AUC) / Receiver-Operator Characteristic (ROC)	**7**
Confusion Matrix	**2**
R-squared R^2	**4**
Root Mean Squared Error (RMSE)	**2**
Standard Root Mean Squared Error (SRMSE)	**1**
Deviance Information Criterion (DIC)	**1**

2.5 Current challenges in DF spatiotemporal modelling research

Upon reviewing the previous works on DF spatial and spatiotemporal modelling, several challenges were identified as follows:

1. Despite previous spatial and spatiotemporal modelling studies, developing a suitable comprehensive methodology to improve spatiotemporal modelling performance is critical.
2. Complete data acquisition using local resources is another challenging task and remains difficult for research in developing countries.
3. Based on the reviewed studies, traditional statistical approaches are commonly used for modelling various aspects of DF. However, integrating GIS techniques with advanced machine learning is essential for better modelling performance.
4. Most studies do not address the issue of data quality and its impact on the prediction model's performance.
5. It is essential, but difficult, to use knowledge from previous research to understand the DF spatiotemporal patterns in the presence of MD.

6. Very few of the previous studies have compared the various approaches and models used to predict DF in terms of spatiotemporal factors, and measured the accuracy of those models.
7. Insufficient utilization of simulation tools following proper analysis at various modelling stages for a better understanding of the DF disease spatiotemporal patterns.

2.6 Summary

Various statistical and ML approaches to spatial and spatiotemporal modelling of DF disease, along with the risk factors affecting the DF disease, have been reviewed. Although most of the papers adopted statistical approaches to model DF transmission, ML methods appear to provide better prediction in the modelling of DF transmission compared to the traditional statistical analysis. Moreover, ML is more appropriate for big data and complex models. Dengue transmission modelling on spatial and spatiotemporal scales is complex since it is influenced by various variables. Therefore, four main categories were established in order to classify the related risk predictors in each model to cover climatic, demographic and socio-economic, entomological, and environmental factors. Temperature, rainfall and population density were the most important of these factors. In general, most modelling approaches in the reviewed papers used regression methods to analyse and model the transmission of disease, the most popular being the GWR and OLS. To significantly improve spatial and spatiotemporal modelling and hot spot analysis, the common risk factors in each category should be considered in addition to a combination of statistical and ML approaches. Moreover, MD is a significant issue that needs to be tackled as it can affect the findings of the model and produce biased estimates. Thus, developing a comprehensive framework will be useful for other researchers undertaking the modelling process as they take into account all

implementation steps from the pre-processing stage to the model accuracy measurements. Ultimately, it offers a modelling framework based on high-quality data and various approaches to improve the spatiotemporal modelling performance.

CHAPTER 3

DATA AND METHODOLOGY

3.1 Introduction

This chapter describes the methodologies applied in this thesis to achieve the stated objectives. The overall methodology and the associated algorithms are explained, and the study area and its specific characteristics are discussed. Data sources and the data acquisition process are described in detail, followed by ethical considerations. Then, the implementation of the methodologies is discussed. After that, the significance of the considered parameters are explained. Finally, the tools used to implement the methodologies are described, followed by the chapter summary.

3.2 Overall methodology

In this research, the principal factors associated with DF spatiotemporal patterns were examined, and previous modelling approaches were investigated in order to develop a comprehensive framework to improve dengue fever (DF) spatiotemporal prediction modelling. To achieve the thesis objectives, the proposed methods involve the different stages are shown in Figure 3.1. The first stage following the data acquisition is the pre-processing stage which involves the collection of data on the incidence of DF in Saudi Arabia, and related climatic, socio-economic, and environmental information. In the second stage, a specific algorithm is used to estimate the missing data (MD) values. Then, the collected data is clustered and categorized based on feature similarities. After that, additional analysis is performed to improve the prediction performance of DF spatiotemporal models. In the final stage, a simulation approach is adopted to predict the main factors that impact DF spatiotemporal patterns, and to identify the risk areas within

the study area based on the historical recorded data. The methodology involves the application of several algorithms: Self-Organizing Maps (SOM), Density-Based Spatial Clustering of Applications with Noise (DBSCAN), Ordinary Least Square (OLS), and Geographically Weighted Regression (GWR). The machine learning (ML) approaches used to classify confirmed cases were: Decision Tree, K-Nearest Neighbours, Random Forest, AdaBoost, Support Vector Classification (SVC), CatBoost, and Naive Bayes. A single simulation approach using the Cellular Automata method was adopted to simulate the DF cases in order to predict future threats. Several tools were used to conduct the analysis: GIS (ArcMap and QGIS), Python Programming Language, and Remote Sensing data (e.g., MCD12Q1). More details regarding the methodology are given in the subsections below.

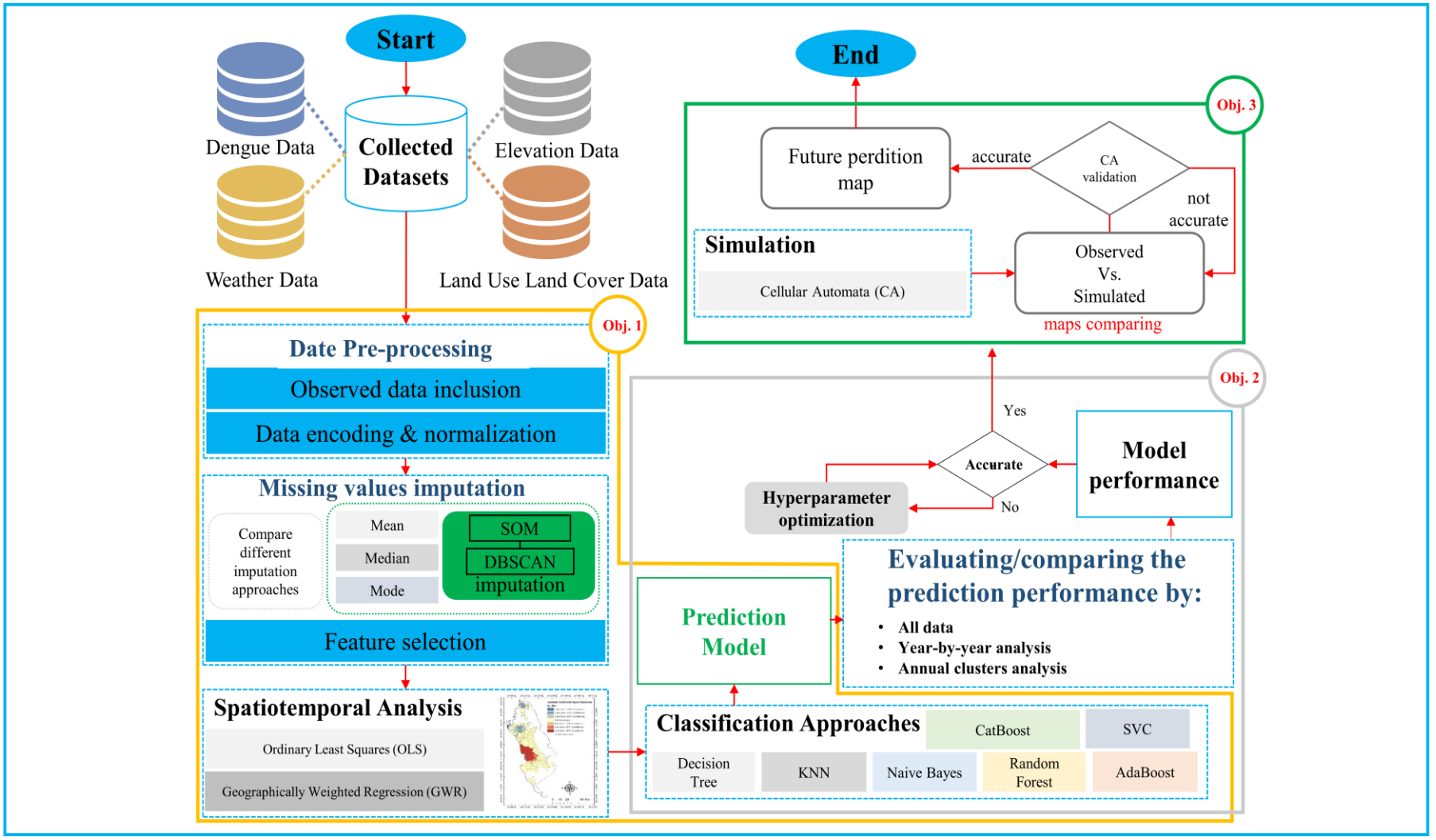


Figure 3.1. Overall methodological flowchart for dengue fever spatiotemporal modelling

3.3 Study area

Jeddah city

Jeddah city (21.4858° N, 39.1925° E) was chosen for this study (Figure 3.2). It is located in the western part of Saudi Arabia in the Hejaz region, and it is the Kingdom's most important port and business centre, dubbed the "commercial capital of Saudi Arabia" (Murad 2007). According to Saudi's general statistical authority, the city's population was 3,457,794 in 2021, which represents 14% of the total population of Saudi Arabia (Murad 2007). The city has a residential density of 2,500 people per square kilometre (Aljoufie and Tiwari 2021). Jeddah is the second largest city in the Kingdom, after the capital city of Riyadh. Another important aspect of this city is that it serves as the main gateway to Mecca for Muslims worldwide who are fulfilling their religious obligations. Mecca is approximately 65 kilometres from Jeddah, while Medina located 360 km north, is the second most holy city (Al-Raddadi et al. 2019; Ayyub et al. 2006). Jeddah is a coastal city by the Red sea and has coastal weather; the relative humidity in Jeddah is usually high most days of the year, particularly in the summer. During the winter, the relative humidity remains quite low at only 15%, although the maximum relative humidity can reach over 95% (Alkhaldy 2017). While this city has an average annual temperature of 28.7°C, it varies throughout the year from 24.22 °C to 32.44 °C depending on the season (Hashem et al. 2018). Rainfall varies significantly depending on the season. Between November and April, Jeddah receives 47.8 millimetres of rainfall, and from June to September it receives 0.5 millimetres (Hashem et al. 2018). DF is most commonly found in Jeddah city based on documented cases (Alkhaldy 2017). Thus, all analyses took into account the total area of the city, considering the district as the spatial scale.

The importance of selecting Jeddah city as the study area lies in:

- The most important port and business centre for the country;
- high residential density;
- the main gateway to the Muslim holy cities (Mecca and Al-Medina);
- suitable weather for dengue vectors;
- the country's most reported DF cases and the most well-documented.

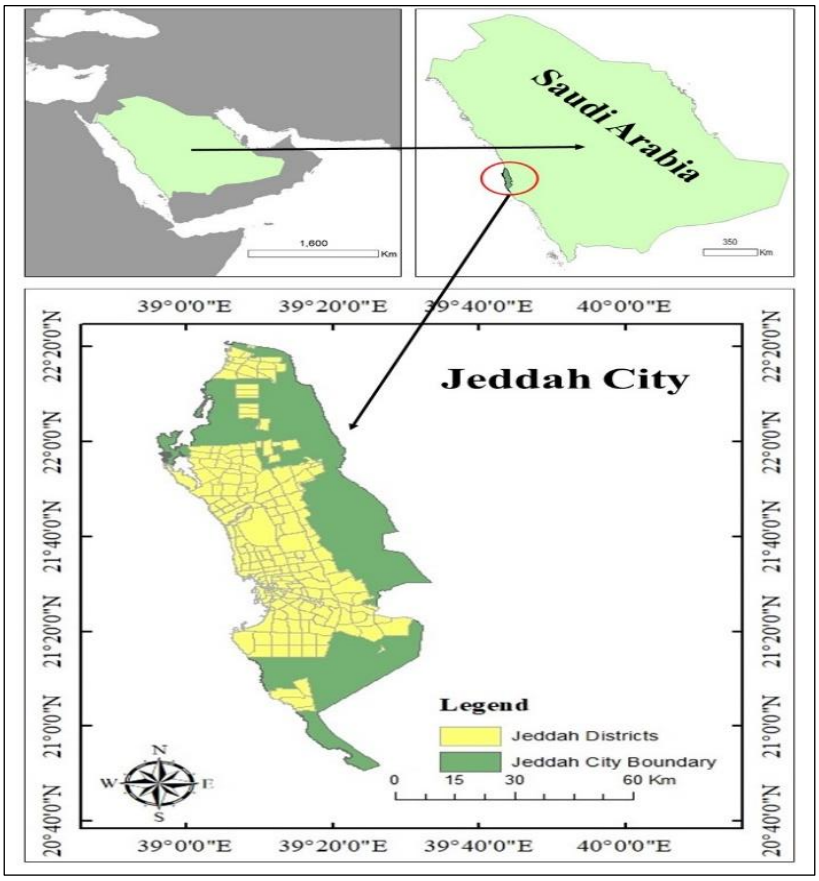


Figure 3.2. Study area: Jeddah city, Saudi Arabia

3.4 Data collection

In this study, the datasets for DF incidents were collected from the Jeddah Health Affairs department after obtaining ethical approval from the relevant authorities (University of Technology Sydney and Saudi Arabia Ministry of Health) as explained in

Section 3.5. The data comprises demographic information about each case: patient gender, age, nationality, registration date, hospital name, status of the incident ("confirmed or suspected"), and district. The collected data provided all the main socio-economic information required for this research. Unfortunately, other socio-economic data related to Gross Domestic Product (GDP), housing, education, and water storage methods are unavailable. Table 3.1 shows the data sources, spatial scale, spatial resolution, temporal scale, and time period.

Table 3.1. Source of data used

Data	Spatial Scale	Satellite Imagery Resolution	Spatial resolution	Temporal Scale	Time Period	Source
Dengue Fever Cases	Local	-	-	Daily (individual cases)	2012 - 2018	Ministry of Health (MOH), Saudi Arabia
Population	Global	1 km (30° × 30°)	1 km	Annual	2012 - 2018	LandScan
Elevation	Global	1 arc second	30 metres	-	2000	SRTM 1 Arc-Second Global from United States Geological Survey (USGS)
LULC	Global	300 metres	300 metres	-	2015	The Land Cover CCI Climate Research Data Package (CRDP) by Space Agency (ESA)
	Global	500 metres	500 metres	Annual	2012 - 2018	MCD12Q1 - USGS
Climatic						
Temperature and Humidity	Global	0.2°	~ 38 km	Annual and Monthly	2012 - 2018	The National Centres for Environmental Prediction (NCEP) Climate Forecast System (CFS)
Precipitation	Global	0.05°	-	Annual and Monthly	2012 - 2018	Climate Hazards Center US Santa Barbara
Wind Speed	Global	0.1°	~ 4 km	Annual and Monthly	2012 - 2018	NASA's Earth Science program

The use of satellite imagery in epidemiological studies enables the identification of important environmental variables that influence the patterns of vectors as well as their interactions (Espinosa et al. 2016). Thus, all data were collected using satellite images except for disease data collected from the Saudi government's database as discussed below.

Dengue cases data: Dengue cases reported by hospitals and health centres in each city are monitored and documented by the Department of Vector-borne Diseases and Zoonotic of the Health Affairs Department of the Saudi Ministry of Health. Records are saved in an Excel file containing the following data: investigation ID/patient code, an investigation created date/time, age, gender, nationality, occupation, initial district, district, initial neighbourhood code, neighbourhood code, initial address, address, the date the symptoms appeared, hospital name, whether the patient was hospitalised ("YES"), whether the patient was not hospitalised ("NO"), the final diagnosis, international epidemiological week, and the date when the blood sample was taken. Additional variables were extracted and added to previously-obtained data, such as the investigation day, week, month, year, municipality, the district under which neighbourhoods were grouped, and city name since some district names are the same in different cities. Some attributes were discarded as explained later in the feature selection and reduction section. The historical data on DF for the period from 2012 to 2018 was collected.

Climatic data: By means of the Google Earth Engine, it is relatively easy to access huge amounts of satellite and weather data. Moreover, with the significantly improved processing capability of parallel computer resources, satellite imaging has become a feasible means of delivering yield projections across wider areas in near real-time (Gorelick et al. 2017; Schwalbert et al. 2020). Hence, all climatic features of interest in the study area were collected monthly and annually from 2012 to 2018 from the Google Earth Engine (GEE) using the JavaScript export function. The land surface temperature and humidity were obtained from the Research Data Archive at the National Center for Atmospheric Research (NCAR) (Saha et al. 2010; Saha et al. 2011). Land surface temperature images produced by remote sensing at moderate spatial resolution have a

smaller spatial scale and provide a more accurate representation of environmental conditions (Yue et al. 2018). The precipitation data were collected from Climate Hazards Group InfraRed Precipitation with Station (CHIRPS) (Funk et al. 2015). Lastly, the wind speed data was collected from the National Center for Environmental Information (NCEI) (Baker-Yeboah and Kilpatrick 2016).

Population data: ORNL's LandScan is an open-source community standard that provides annual statistics for worldwide population distribution and all population data were obtained from it. The database is updated annually, with the latest statistics made available in the third quarter of each calendar year. It is designed to separate census counts within a region boundary by utilising the best available demographic (census) and geographic data, remote sensing, and image-processing tools within a multivariate dasymetric modelling framework (Bright et al. 2013, 2014, 2015, 2016; Bright et al. 2017; Rose et al. 2018; Rose et al. 2019).

Elevation and Land use land cover data: Environment-related data can be collected from several providers of satellite images. For this research, elevation images were downloaded using Shuttle Radar Topographic Mission (SRTM) 1 Arc-Second Global from the United States Geological Survey (USGS) (Earth Resources Observation Science Center 2018). Because mosquito reproduction is influenced by the type of land cover (Wiese et al. 2019), several researchers have used LULC or satellite-derived indices as proxies for appropriate regions of dengue vector to model the disease (Tran et al. 2020). Thus, LULC was obtained from two different sources to investigate the influence of LULC on the disease. First, the global data regarding land usage and land cover was obtained from the European Space Agency (ESA) Climate Change Initiative (CCI)

(Defourny et al. 2017), which can provide data spanning 1992 to 2015. These data are available at 300m spatial resolution and provide multiple land factors such as agricultural "rainfed" land, herbaceous cover, tree cover, needle-leaved, deciduous, sparse vegetation (tree, shrub, and herbaceous cover), urban areas, bare areas, consolidated bare areas, unconsolidated bare areas, and water bodies. Second, in order to determine whether there were any changes in land cover during the period of interest, annual land cover images were extracted from the online MCD12Q1 dataset provided by the USGS Land Processes Distributed Active Archive Center (LPDAAC) using the NASA Earthdata Search tool (https://lpdaac.usgs.gov/products/mcd12q1v006/, last accessed on February 22, 2022). The extracted LULC features include open shrublands, grasslands, urban and built-up areas, barren land, and bodies of water (Friedl 2019).

3.5 Ethics approval

The data collection procedures were approved by the Human Research Ethics Committee, University of Technology Sydney (UTS) application ID (ETH 194366), and by the Research and Studies Department - Jeddah Health Affairs representative of the Saudi Ministry of Health (Research number: 01203).

3.6 Implementation of the methodology

3.6.1 Descriptive analysis

The original dataset contained 37,903 confirmed and suspected cases recorded from the first of January 2012 to 31 December 2018. The patient records contain demographic information such as age, gender, occupation, hospital name, hospital admission data, district, address, and the case diagnosis. This data was used to calculate the number of

confirmed cases in each district; a descriptive analysis was conducted based on these parameters.

3.6.2 Data pre-processing

Anomalies, trends, and correlations in big data sets can be discovered through the data mining process (Wu et al. 2013). The mining involves various techniques such as association rule learning, regression, summarization, classification, and clustering (Nisha et al. 2013). It is a critical issue requiring a great deal of engineering effort to reduce bias (Sessa and Syed 2016). The level of potential bias may depend on the reasons for the missing values and the procedures applied to correct the MD problem. Therefore, data analysis containing missing values necessitates meticulous planning and attention (Jakobsen et al. 2017). Moreover, the data mining process and the quality of the data have a significant influence on the modelling performance (Kamkhad et al. 2016; Sessa and Syed 2016).

Usually, data preparation is essential in most research studies (Alshehri 2019). This stage includes data cleaning methods such as removing noise and irregular data. Data transformation is the process of transforming and consolidating data so that ML technologies can use it. Feature aggregation and data exclusion involve the selection and extraction of features. Data cleaning is the process of correcting standardising data, screening out untrustworthy data, and "excluding" extra aspects of the data. A data normalisation process was used to guarantee that all attributes were displayed using the same measurement units and range. Data normalisation is intended to give the same weight to all characteristics, which is beneficial in statistical learning approaches. In this current study, several steps were taken to improve the data quality. Each of these steps is described below.

Data Cleaning: Real-world datasets acquired over a period of time are often incomplete and inconsistent, and may contain noisy data, resulting in database inconsistency (Al-Hagery et al. 2019). Patient data was initially recorded in the Arabic language. During the first stage of data analysis, it was found that the data entries were not standardized and contained many spelling errors and abbreviations. For instance, the gender parameter of "Male" and "Female" expressed in two languages (Arabic and English), had spelling errors, was marked as unknown or contained an empty cell. Many of the values for the age parameter were expressed in numbers or letters, or were missing or declared unknown. The names of the districts presented the most significant challenge as they contained many spelling errors, or the areas were given local names, not those assigned by the municipality of the area. At this stage, the data that needed to be amended during the data transformation stage were identified, and irrelevant data was excluded.

Data integration: The patient data were obtained from an Excel file; the climatic, population, elevation, and land use/land cover data were obtained from numerous satellite images. To have a consistent format, yearly and monthly imagery for the collected parameters were aggregated using the ArcGIS 10.4 Spatial Analyst Tool Cell Statistics. The “Zonal Statistics as Table” aggregates the variables in the images according to the annual and monthly total, average, maximum, and minimum into every cell of the target grid within the district polygon using GCS-WGS-1984 Geographic Coordinate Systems. Yearly and monthly averages were calculated for the total population, temperatures, humidity, and wind speed; for the precipitation variable, the total amount of precipitation per month and year were considered for each district of Jeddah city. Although cell statistics are not required for elevation and land use land cover (LULC) images, zonal statistics are required to calculate these variables' values for all the study area districts.

The unique identification code provided by the zonal statistics table tool was used to combine extracted values with the study Shapefile. Moreover, these data were further investigated to determine their relationship with the incidence of DF. Figure 3.3 shows the pixel size and polygon to illustrate the mechanism of calculating desired parameters per district. Moreover, for spatial analysis, patient data were aggregated into the same Shapefile for each district "polygon" by counting the confirmed cases, suspected cases, male, female, Saudi, and non-Saudi. Later, the Shapefile data was exported to the Excel file and linked to the patient file for further analysis, MD imputation, and spatiotemporal prediction modelling.

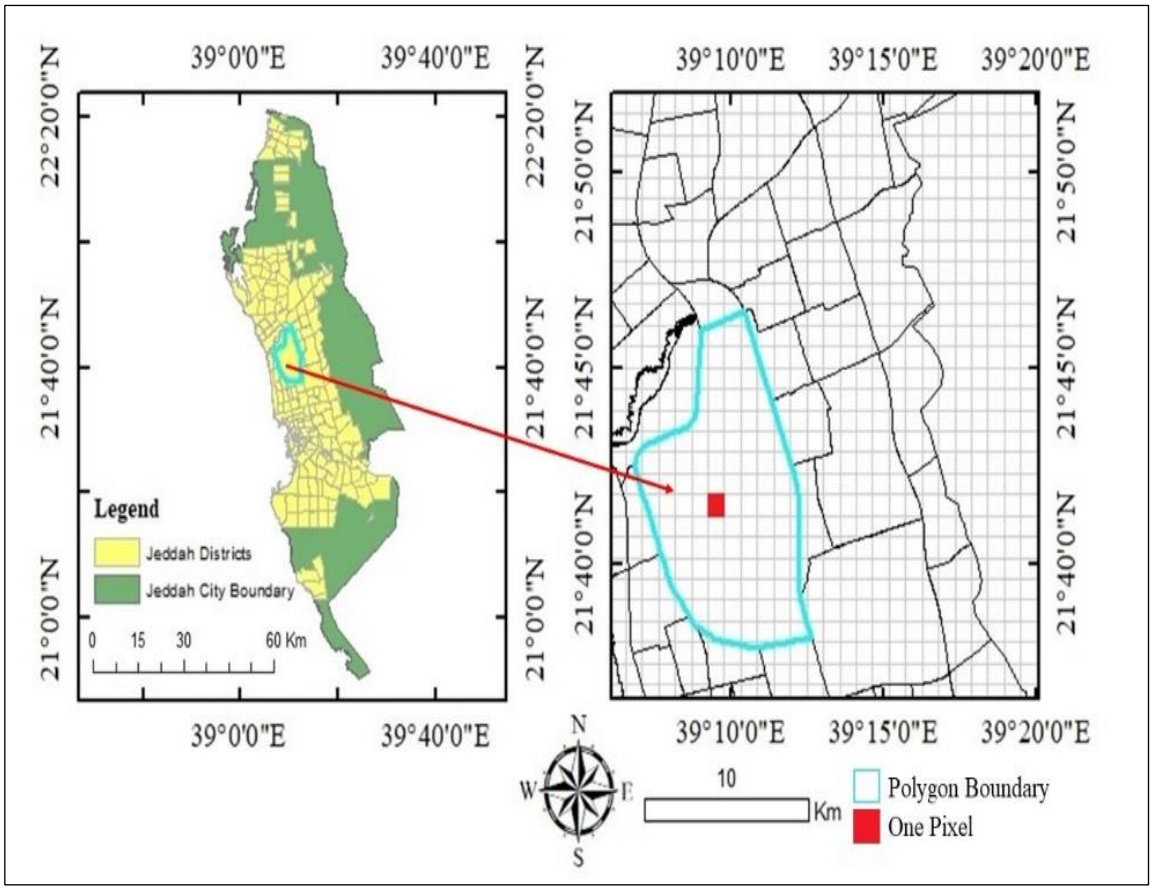


Figure 3.3. Example of dividing cells and calculating features within a single district

Data transformation: In this stage, the data were unified and translated manually into English using Microsoft Excel 2016 for further analysis. The age data were converted to numerical form; for instance, 12 months old was changed to 1 (i.e., one year). Gender was assigned either to “Male” or “Female”, while cells with “unknown” values were considered as missing values. Nationality was expressed uniformly as the country's name; for example, ‘Saudi’ was recorded as ‘Saudi Arabia’. Later, another copy of the data was created, and nationalities were recorded as either Saudi or non-Saudi for further analysis.

Data inclusion/exclusion: This stage involved all parameters collected through the satellite images and computed based on each district. However, all patients’ records with the missing dependent parameter of the final diagnosis were excluded. The occupation feature was excluded as it was partially available for only four years in the dataset. Moreover, several attributes were discarded for various reasons:

- The attributes were not related to the study objectives.
- Most of the values were missing and could not be imputed.
- Duplicated parameters.

The unrelated parameters included “was the patient hospitalised? YES/NO”, “the date the symptoms appeared”, and “sampling date”. In addition to previous unrelated parameters; “was the patient hospitalised? YES/NO” and “sampling date” were unavailable for most of the years. For the second exclusion criteria, “Initial Address” and “Initial District Code” were not available for all the years, while “occupation” data were available for only four years, and most of the records had missing values. Therefore, these three parameters were discarded. In terms of duplicated parameters, “Initial District Name”, “Initial District Code”, “initial address”, and “address” were discarded and

grouped under "original district". The "Week no." row contains the week of the year when the case was recorded. Table 3.2 shows the data for reported dengue cases and for missing records for each year.

Table 3.2. Features containing missing values in the obtained data

	2012	2013	2014	2015	2016	2017	2018
Investigation Id	✓	✓	✓	✓	✓	✓	✓
Week No.	✓	✓	✓	✓	✓	✓	✓
Investigation Created Date/Time	✓	✓	%	%	✓	%	%
Age	%	%	%	%	%	%	%
Gender	%	%	✓	✓	%	%	%
Nationality	%	%	%	%	%	%	%
Occupation	%	%	%	%	✗	✗	✗
Original District	%	%	%	%	%	%	%
Initial District Name	✗	%	✗	✗	✗	✗	✗
Initial District Code	✗	✗	✗	✗	✗	✗	✗
District Code	%	%	✗	✗	✗	✗	✗
Initial Address	✗	✗	✗	✗	✗	✗	✗
Address	%	✗	✗	✗	✗	✗	✗
The Date the Symptoms Appeared	%	%	%	%	%	✗	✗
Hospital Name	✓	✗	✓	%	%	%	%
Was the Patient Hospitalised: YES	%	%	✗	✗	✗	✗	✗
Was the Patient Hospitalised: NO	%	%	✗	✗	✗	✗	✗
The Final Diagnosis	%	✓	✓	%	%	%	✓
International Epidemiological Weeks	✓	✓	✗	✗	✗	✗	✗
Sampling Date	%	✗	✗	✗	✗	✗	✗

* ✓ = all records are available, % = some records are missing; ✗ = all records are missing

Normalization: The collected data contained a mix of numerical and categorical variables. Hence, the categorical values needed spatial treatment before input into the predictive model. Therefore, binary encoding was applied to these features to convert them to binary numerical values to fit the desired model. After converting all the data to numerical values, a normalization method named "StandardScaler" from the Sklearn library was adopted and applied to scale all values of features to fit within a range of 0 to 1 (Pedregosa et al. 2011). This stage is essential to reduce potential bias that may affect the results.

MD imputation: This study adopted a novel strategy to replace the missing values of spatial and temporal features, with the most common values in patient clusters found by applying the SOFM and DBSCAN approaches. Further details are given in the section discussing the methodologies applied to achieve the first objective.

3.6.3 Objective 1 (Develop a data analytical model in the presence of MD)

Figure 3.4 depicts the overall methodological approach used to achieve the first objective, and compares it to earlier state-of-the-art approaches. Following the data collection and pre-processing stage explained earlier, two ML approaches were adopted (SOFM and DBSCAN) to create clusters of patients sharing similar features, and impute missing values. Any missing features in a particular cluster were given the most frequent values of those features within that cluster. After preparing the data and determining the inclusion parameters, further analysis was conducted, including spatiotemporal analysis, to determine the risk areas and classify patient records.

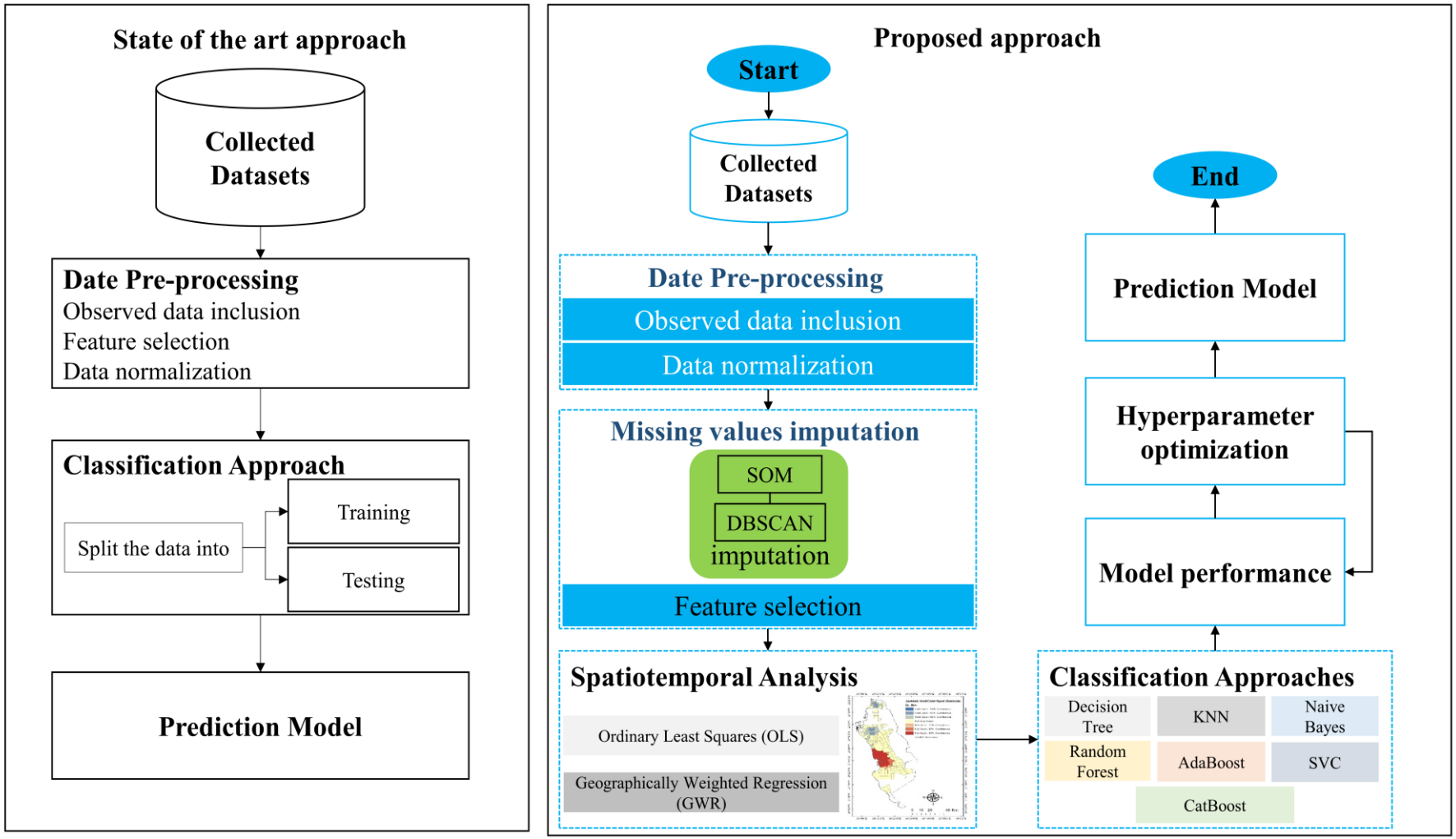


Figure 3.4. First objective: methodological flowchart for DF spatiotemporal modelling

3.6.3.1 Self-Organizing Feature Map (SOFM)

A Self-Organizing Feature Map (SOFM) or Self-Organizing Map (SOM) is a commonly-used unsupervised neural network based on competitive learning (Germano 1999). It was introduced initially by Kohonen in 1982, and is also known as the Kohonen map (Mutheneni et al. 2018; Valles et al. 2019). The SOFM has been applied recently in numerous engineering, health, and geospatial scenarios to classify different problems. The SOFM is an effective tool for clustering, reducing the data dimensions, and visualizing the data patterns (Germano 1999; Mutheneni et al. 2018). The integration of SOFM and spatial GIS applications offers a novel means of analysis (Mutheneni et al. 2018). Here, the output of SOFM helps in performing the clustering of cases with similar features which in turn will improve the data input. Vesanto (Vesanto and Alhoniemi 2000) has explained the steps involved in the SOFM clustering algorithm. Generally, this approach required several parameters which are: the specific parameter to control the learning rate per neighbour "std", learning rate "$step$", learning reduction parameter "$reduce\_radius\_after$", reduction rate for step parameter "$reduce\_step\_after$", and reduction rate for std parameter "$reduce\_std\_after$". In this study, the SOFM method was adopted to map the data according to two dimensions in order to place similar features into one group, similar to (Shukla et al. 2018). Although SOFM provides clustering patterns (Mutheneni et al. 2018), this method cannot provide clusters, and clustering approaches such as k-mean or DBSCAN are needed to extract clusters from data (Shukla et al. 2018). Lastly, SOFM is a useful tool for maintaining the key topological properties of the data (Valles et al. 2019). Moreover, it has been shown that SOFM is the appropriate method to use when a dataset contains missing values as it can estimate the missing values in a particular record by using the centre of each subclass (Latif et al. 2010).

3.6.3.2 Density-based spatial clustering of applications with noise (DBSCAN)

Clustering is an important descriptive method used in data mining. It organizes the data into meaningful classes or clusters, with objects that are related to one another placed inside the same cluster, but different from those in other clusters (Vesanto and Alhoniemi 2000). For spatial-based datasets, class identification tasks are generally performed by using clustering approaches. Moreover, of the six different clustering methods, the density-based approach is the most appropriate for clustering spatial data (Vesanto and Alhoniemi 2000). Nafees (Ahmed and Razak 2014) illustrated the advantages of DBSCAN, one of which is that it does not require the cluster numbers to be predetermined, as this is done automatically. Second, it requires the user to establish only two parameters: the value of the radius of neighbourhood "eps", and the minimum number of points in each cluster "$MinPts$". Third, it can discover any cluster shapes. Lastly, it can effectively distinguish noise points and is robust against outliers. Hence, the advantages of DBSCAN make it appropriate for the current study where patients' clusters are extracted from SOFM data. Lastly, each cluster extracted by DBSCAN is used to impute the MD with the "mode" to create the desired model.

After completing the pre-processing stages and filling in MD, several stages are implemented before the spatiotemporal modelling to determine the included features in the prediction model. Also, the data is split into training and testing datasets to validate the model's performance as explained below.

Correlation analysis/Feature selection: A correlation analysis was used to determine the strength of the relationship between the dependent variable (number of confirmed dengue cases) and other time-series independent variables (climate and population variables). Demographic features were excluded at this stage as they are not suitable for

the desired analysis; the elevation and land cover parameters were also excluded since they can change slightly over time (Ashby et al. 2017). Moreover, there is no direct relationship between vector-borne disease and land cover (Jácome et al. 2019). The coefficient correlation values range from -1 and +1, indicating a positive correlation when the value is close to +1 and a negative correlation when the value is close to -1. Then, positively-correlated features were included in the prediction models, and the negatively-correlated parameters were excluded from the dataset.

Cross-validation: Cross-validation was conducted at this stage to evaluate the results of the training and testing sets. Time series split cross-validation was applied to determine the model's accuracy; 70% of reported DF cases were used as the training dataset and the remaining 30% of the data were used for the testing.

3.6.3.3 Spatiotemporal analysis and GIS mapping

In order to control the transmission of DF, it is crucial that the spatial and temporal patterns of the disease be investigated. Therefore, in this study, ArcMap 10.4 software was used to depict the spatial patterns of dengue disease in Jeddah's 205 districts, together with the data collected for climatic, demographics and environmental factors at one year and as the temporal pattern. In each district, the cumulative incidence of DF, climatic, demographic, and socio-economic characteristics were converted into polygon Shapefile. Ordinary Least Square (OLS) and Geographically Weighted Regression (GWR) are common approaches used to investigate the spatial relationship between disease and other factors. GWR is better suited for use at the local scale and can handle spatial autocorrelation better than OLS, especially when there is multicollinearity among several parameters (Khormi and Kumar 2011; Mutheneni et al. 2018). However, both approaches

are adopted here and compared using the Akaike Information Criterion (AIC), R^2, and adjusted R^2 to determine the most appropriate method for the collected data to establish any associations between the incidence of DF and the causative factors. The value of R^2 indicates a model's capacity to explain a variable's variation; the higher the value, the better the performance of the model is. The AIC measurement is used to indicate the model's accuracy; the smaller the value, the better fitting is the model (Acharya et al. 2018; Tu and Xia 2008).

3.6.3.4 CatBoost classifier

CatBoost is a powerful ML technique that generally uses decision trees as base predictors (Dorogush et al. 2018; Prokhorenkova et al. 2018). Moreover, it can handle databases containing numerical and categorical features for prediction (Ester et al. 1996). Therefore, given the features and heterogeneity of the dataset, the CatBoost method was adopted to predict DF spatiotemporal patterns. Following the previous stage which extracted patient clusters using SOFM and DBSCAN, the CatBoost classifier was applied, and the model accuracy was compared with that of other approaches namely, the Decision Tree, k-Nearest Neighbours (KNN), Random Forest, AdaBoost, Support Vector Classification (SVC), and Naive Bayes.

After extracting and imputing MD, the final dataset was split into "training" and "testing" sets. In the split, a 70:30 ratio was maintained using the sklearn "sklearn.model_selection.train_test_split" method (Pedregosa et al. 2011); 70% of the data were used as training for the model, while the remaining 30% were used to test the model. So as to determine the approach with the best performance, the different modelling approaches were applied and compared. Then, a confusion matrix was used to illustrate

the model performance by comparing the actual data with the predicted values, as shown in Table 3.3. The model accuracy was measured with the following equation:

$$Accuracy = (TP + TN)/(TP + TN + FP + FN) \quad \textbf{(3.1)}$$

Table 3.3. Confusion matrix for the prediction models

	Predicted confirmed cases "1"	**Predicted suspected cases "0"**
Actual confirmed cases "1"	True Positive (TP)	False Positive (FP)
Actual suspected cases "0"	False Negative (FN)	True Negative (TN)

3.6.3.5 Model performance

The performance and accuracy of many ML algorithms are determined by the hyperparameter settings and the method used to set the hyperparameters (Cui and Bai 2019; Eggensperger et al. 2013). Several hyperparameter approaches have been adopted previously including Gradient optimization (Chapelle et al. 2002), Bayesian Optimization (Snoek et al. 2012), and Random search (Bergstra and Bengio 2012). Moreover, for decades, Grid search has been the de facto parameter optimisation standard in machine learning, consisting of an exhaustive search of a manually-selected subset of the learning algorithm's hyperparameter space. Grid search is the state-of-the-art method applied for hyperparameter optimisation due to its simplicity of execution and parallelization, as well as its robustness in low-dimensional spaces (Belete and Huchaiah 2021). Therefore, the current study conducted an exhaustive grid search to tune the hyperparameters of an estimator through sklearn to assess the model's performance (Pedregosa et al. 2011). In the training stage, the grid search parameters were set to a range of values involving the default values in addition to values that led to good performance in previous studies.

Following achieving the first objective, the second objective is about improve the finding of 1st objective as explained in coming section.

3.6.4 Objective 2 (Improve MD imputation and prediction model performance)

3.6.4.1 MD imputation methods

This section describes the methodologies utilised to impute values in the original incomplete dataset, and explains the ML-based prediction algorithm. MD is a critical issue in the pre-processing stage, and several approaches can be used to handle it according to the type and quantity of missing information. However, the imputation technique is the usual approach taken to address the missing values problem due to its meaningful replacement (Sessa and Syed 2016). Several traditional imputation techniques, including mean, median, and mode, have been applied to solve the problem and obtain the missing values. Generally, these approaches replace the missing values for a particular feature with a value (mean, median, and mode) found in that attribute. Although missing values in the collected data are more than 24.7% for the district feature, thereby exceeding the 5% limit for discarding records (Jakobsen et al. 2017), one common approach was applied: to discard the records containing missing values in order to assess the prediction model's performance using complete records only. This strategy has several variations that consider the amount of MD and remove instances or characteristics with a high level of MD. When discarding a characteristic, it is vital to establish the characteristic's relevance to the remainder of the data. Similarly, before discarding an instance with missing values, the size of the collection must be considered. Frequently, the discarding of relevant characteristics is not recommended, even if they have a high number of missing values, and the same applies to a small dataset (Sessa and Syed 2016).

The Self Organizing Feature Map (SOFM) has the advantage of being able to analyse and understand the complex characteristics of high-dimensional data. SOFM maps the data into a basic low-dimensional representation, making it easier to see the data's complexity (Faisal et al. 2010; Valles et al. 2019). The SOFM was utilised in dengue disease studies to identify the principal risk factors that can distinguish dengue patients (Faisal et al. 2008). Srinivasa (Mutheneni et al. 2018) used it to identify endemicity patterns in specific locations. Moreover, the SOM method is appropriate for the classification and imputation of data with absent values as it reduces the error rate in the classification (Sommer et al. 2003). Here, an advanced imputation algorithm was adopted that combines the advantages of the Self-Organizing Feature Map (SOFM) and Density-Based Spatial Clustering of Applications with Noise (DBSCAN) to impute MD; the results obtained by this algorithm were compared with those obtained using traditional imputation approaches (mean, median, and mode). SOFM is an unsupervised learning technique that is generally suitable for clustering tasks, reducing data dominations to a lower dimension and enabling the visualization of complex data (Shukla et al. 2018). The main advantage of this method is its ability to maintain the data relationship when transforming a complex data dimension into a simpler low-dimension (Asan and Ercan 2012). The SOFM algorithm is based on unsupervised learning where the training is fully stochastic and based on data, and does not require information about input data (Sorjamaa et al. 2007). Although the SOFM method assists with the formation of clusters, it cannot detect and extract the clusters relying on cluster algorithms such as k-mean or DBSCAN to do so (Shukla et al. 2018). In this study, the DBSCAN approach was adopted due to its ability to locate clusters with any shape and noise points, and it has been expanded to handle a wide range of data formats including spatiotemporal data (Shi and Pun-Cheng 2019). Lastly, after identifying and extracting the clusters, three imputation algorithms

were applied (mean, median, and mode imputation) to each cluster and each approach was compared to determine the one that best fit the data based on variance value: the lower the variance value produced by each algorithm, the better the fit.

3.6.4.2 Build prediction models

In this step, several prediction models were applied to predict the probable occurrence of confirmed cases based on the positively-correlated variables. These models are Decision Tree, Random Forest, K-Nearest Neighbours (KNN), AdaBoost, Support Vector Classification (SVC), CatBoost, and Naive Bayes classifier. All the models were adopted and implemented in Python programming language using functions available from the sklearn library (Pedregosa et al. 2011).

As an essential step in developing spatiotemporal prediction models, the collected factors need to be unified in terms of spatial and temporal scales to fit the prediction model. Thus, in this study, one district was set as the spatial scale and one year as the temporal scale based on the historical distribution of the confirmed dengue cases. All of the adopted prediction approaches require the following steps:

1. Unify the data scales by choosing one district and one year as spatial and temporal scales respectively.
2. Propose three different scenarios to improve the model's accuracy.
3. Split the data into 70% and 30% for model training and testing, respectively.
4. Select features by determining strongly-related factors through correlation analysis.
5. Build the classification prediction model based on confirmed cases as the dependent variable, and select features obtained in step 4 as explanatory variables.

6. Create a confusion matrix to illustrate the best prediction model by comparing the performance achieved on the testing dataset.
7. Repeat steps 3-5 for all created scenarios.
8. Compare the prediction model's performance on the test data in all scenarios to determine the optimal approach.

3.6.4.3 Performance evaluation

Several approaches can be used to handle data with missing values depending on the type and number of missing values as mentioned in the literature. In the current study, several imputation approaches were adopted, and their performance was compared with several prediction classifier models. The adopted models are Decision Tree, Random Forest, K-Nearest Neighbours (KNN), AdaBoost, Support Vector Classification (SVC), CatBoost, and Naive Bayes classifier. To assess the accuracy of the adopted models in terms of their approaches to data imputation, three scenarios were proposed to verify the efficiency of the adopted models and determine whether the size of the data had an impact on the performance of the model. In one scenario, the analysis was conducted on the entire body of data; in the other scenario, annual data were analysed. During the third evaluation, the modelling was performed for each cluster using yearly data. For example, applying the proposed imputation methods (SOFM-DBSCAN) using the records for the year 2012 generated seven clusters. Then, the modelling methods were applied to each cluster individually to improve the modelling performance. Lastly, in addition to the models' default parameters, hyperparameters based on grid search were adopted to determine whether they improved the final models. Figure 3.5 illustrates the methodological flowchart for the second objective.

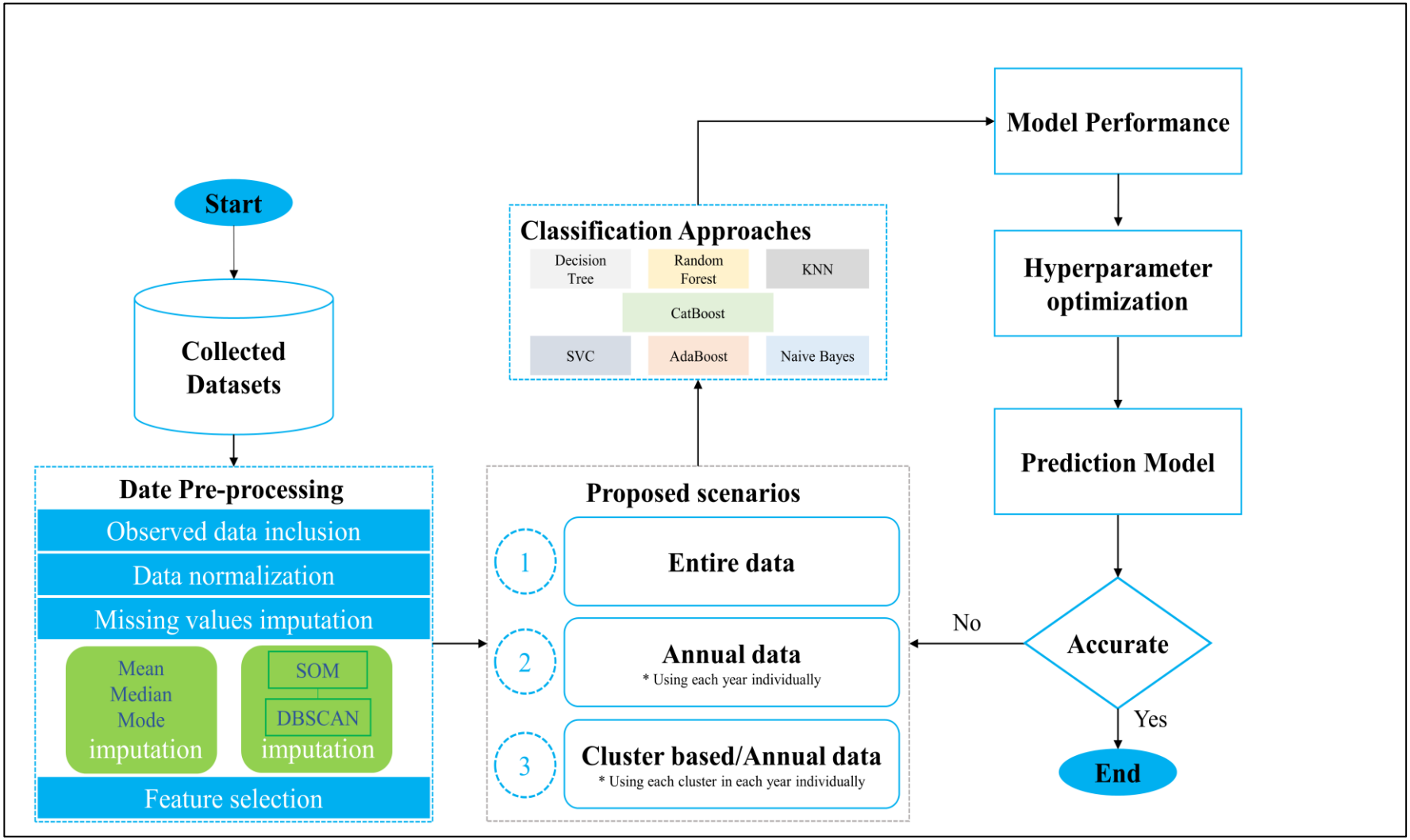


Figure 3.5. Methodological flowchart of the improved model for the second objective

3.6.5 Objective 3 (Simulate risk areas)

Apart from the dengue case data obtained from the Saudi Ministry of Health, this study makes extensive use of satellite images obtained from numerous secondary sources. Thus, to meet the current objective, additional pre-processing for the collected data was performed. Following data preparation, the simulation adopted algorithm is discussed in Subsection 3.6.5.2. The study's methodological flow chart is depicted in Figure 3.6, the techniques and adopted algorithms are discussed in depth in the following sections.

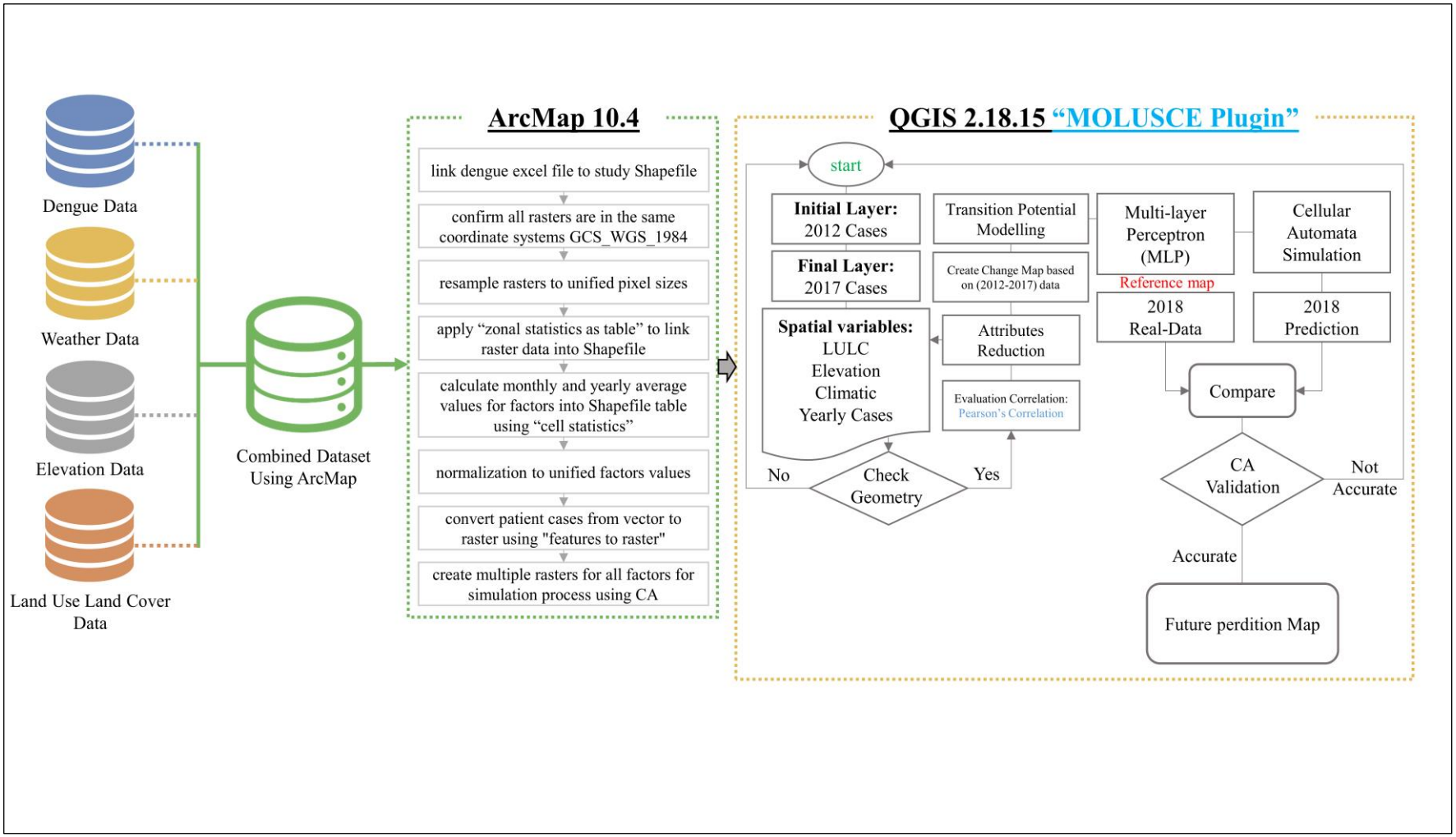


Figure 3.6. Methodological flowchart of the dengue spatiotemporal prediction model for the third objective

3.6.5.1 Data Processing

Relevant data were collected from various sources, and because of the multiple formats, the data needed to be aggregated into a consistent format. For instance, cases data were recorded in Excel, while the climatic, environmental, elevation and LULC data were in rasters. ArcMap 10.4 and Microsoft Excel 2016 were used to handle this aspect. Initially, Excel file patient data were aggregated using pivot tables to extract the total annual number of confirmed cases for each district (2012-2018). Then the tables were imported to ArcMap 10.4 to combine them with Shapefile. Seven different layers were created in ArcMap, one for each year of the study period. The total number of cases for each district was calculated using "Cell Statistics", a software spatial analysis tool. New fields were added to the Shapefile for the total number of confirmed cases, suspected cases, male, female, and the average ages, for each district and on a yearly basis. On the other hand, the pre-processing of rasters was different. Fortunately, all raster projections were extracted using the same geographic coordinate systems (GCS_WGS_1984). However, when rasters are being re-projected to unify the coordinates, this can lead to errors because each re-projected grid cell covers an area of the earth's surface that is different from that of the original cells (Kugler et al. 2015). To maintain the original pixel values, taking the climatic raster as a reference image, all rasters were resampled using the nearest neighbour resampling approach (Wondrade et al. 2014). Care needs to be taken to change the pixel sizes correctly when making the raster resolution consistent; failure to do so can result in incorrect values for features and locations (Deng et al. 2007). Then, "Zonal Statistics as Table" and "Cell Statistics" tools were used to aggregate and calculate the collected factors into the Shapefile. The average statistics for total population, temperature, humidity, and wind speed were derived from the values recorded for the year; for the precipitation variable, the total number per year was considered for

each district of Jeddah city. Then, seven raster images were created based on the confirmed disease cases, and associated factors were created. ArcMap 10.4 was used for all the previous steps and maps to prepare the data for the next analysis stage using QGIS software.

3.6.5.2 Cellular automata (CA) approach

As a means of analysing and forecasting evolving patterns, the Cellular Automata-Artificial Neural Network (CA-ANN) model can provide a detailed view of the complexities of a spatial system (Rahman and Rahman 2021). The cellular automata approach is becoming increasingly popular in the modelling and simulation fields as it is simple yet capable of describing complex systems (Ortigoza et al. 2019). A QGIS-based tool called MOLUSCE-plugin (Modules of Land Use Change Evaluation) was recently created to forecast and analyse LULC scenarios. This plugin can conduct simulations using a variety of models, including weights of evidence, logistic regression, artificial neural networks (ANN), and multi-criteria evaluation. It can also build a transition probability matrix using the CA-MC approach (MCE). Moreover, a series of sequential steps are involved in the plugin: input, analysis of area changes, modelling, simulation, and validation. LULC changes are covered in the CA model both statically and dynamically, and the predictions are highly accurate (Kafy et al. 2021). Although previous studies used this tool to predict the LULC and geomorphological changes in an area of interest (Aneesha Satya et al. 2020; Ferdous and Rahman 2019; Jogun et al. 2019; Kafy et al. 2021; Kamaraj and Rangarajan 2022; Rahman and Rahman 2021; Rahman et al. 2017; Ullah et al. 2019), to suit the purposes of the current study, the MOLUSCE plugin was used to simulate future disease cases based on historical confirmed cases for the study period that were extracted as rasters using ArcMap in the previous stage. To

best of the researcher's knowledge, no previous studies applied the tool "MOLUSCE plugin" to predict dengue cases. In this study, three scenarios were proposed in order to confirm the ability of the adopted tool to predict disease future incidents with a high level of accuracy. In the first scenario, annual data were used to predict the risk map for the following year; this map was then compared with the number of cases recorded for the same year. Thus, annual confirmed number of disease cases from 2012 to 2017 were used as inputs to the model in order to generate a simulated disease map by setting the target year for the validation map to assess the model accuracy. Additionally, collected rasters were used as spatial features for the desired year. The initial input layer was 2012 confirmed cases, the final input layer was 2013 confirmed cases, and the spatial variables were population, climatic, environmental, elevation and LULC for 2012. To simulate the 2014 prediction map, the simulation iteration was set to 1 (one) for the first scenario. In the first stages of the plugin, initial year, final year, and spatial variables cannot be inserted unless the geometries of all rasters are matched. The Pearson's correlation coefficient was used to determine whether there existed any correlation between the annual number of cases and the spatial variables. Then the area changes indicated by the initial and final rasters were calculated and used to create a raster for the transition matrix. The transition matrix, showing the percentage of pixels that change from one type to the next, was also generated by the algorithm (Kamaraj and Rangarajan 2022). The MLP-ANN technique was applied to calculate the potential transmission map as it is robust, particularly in terms of linear regression (Jogun 2016; Rahman et al. 2017). If the Kappa value derived from the previous stage meets the assessment standard, the cellular automata simulation method can be applied to predict the number of cases for the following year.

The second scenario was created to predict future risk maps by adjusting the simulation iteration to simulate the 2018 risk map. The 2012 and 2013 maps were used as inputs to simulate 2018 to maintain the same setting and perform the same steps as those for the first scenario. However, the simulation map was set to 5 as the gap between 2012 and 2013 is one year, and the desired year to be simulated is five years after the final layer. After the 2018 map was simulated, the recorded cases raster for 2018 was used as a reference map to assess the accuracy of the simulated map. The model accuracy was assessed using Kappa statistics: overall Kappa, Kappa histogram, and Kappa location and percentage of correctness. According to Altman (1991) (Lakshita and Rahayu 2021), the Kappa value of 0.81-1.00 shows almost perfect agreement, 0.61-0.80 substantial agreement, 0.41-0.60 moderate agreement, 0.21- 0.40 is fair agreement, 0.01 - 0.20 shows slight agreement, and a value < 0 is considered as less than chance agreement.

The third scenario follows the same steps in the first two scenarios, but the difference is using the average for all values from 2012-2016 to predict the 2018 risk map and compare it to the observed 2018 risk map. So the initial layer will be the average of disease cases from 2012-2016, the final layer will be the year 2017 risk map, and lastly the iteration adjusted to one to simulate the 2018 risk map. The same procedure can be implemented to predict a map for future cases, but with more adjustment regarding the number of iterations, and the gap between the first and final rasters. For example, if 2012 is the initial input and 2017 is the final year to be investigated, the simulation can be set to two iterations indicating the prediction for the following ten years as the gap between the two input years is five years which means that the 2027 risk map is predicted because of the temporal difference between 2018 and 2027.

3.7 Factors considered in this study, and their importance

3.7.1.1 Dengue incidents

In this study, DF spatiotemporal prediction models were developed using historical reported cases. Modelling DF spatiotemporal requires an investigation of numerous spatial and temporal factors known to influence the disease. Hence, common risk factors identified in the literature were considered for the study area. Further details of these significant factors are given below.

3.7.1.2 Climatic factors

Temperature "Land surface temperature": It has been proven that temperature affects not only the survival time and habitats of insect-borne disease vectors, but also their replication, maturation, and infectious periods (Wu et al. 2009). Moreover, a previous study found that climatic variables including temperatures are significant predictors when investigating DF cases, pathogens, and vectors (Chen and Hsieh 2012).

Humidity: The incidence of DF may increase due to climate change. It has been suggested that temperature, rainfall, and humidity all play a role in the transmission cycle of dengue viruses and their mosquito vectors due to their environmental sensitivity (Li et al. 2020). *Aedes* mosquitoes are influenced by relative humidity in several ways including oviposition, egg hatching, flight performance, feeding behaviour, and lifespan (Xiang et al. 2017).

Precipitation: This is an important factor, with numerous studies having identified temperature and precipitation as two of the key risk factors associated with DF (Li et al. 2020). In addition, it has been found that the hatching of mosquitoes differs dramatically

depending on precipitation and humidity (Depradine and Lovell 2004). Lastly, as a result of factors such as temperature and precipitation, mosquito vectors and pathogens transmitted by these vectors can differ greatly depending on the local climate (Sippy et al. 2019).

Wind Speed: The incidence of dengue is also affected by wind speed (Fairos et al. 2010). Wind causes dengue vectors to lose their host-seeking flying activity, which eliminates oviposition and human contact (Mala and Jat 2019a). Moreover, dengue vector distribution and oviposition are favoured by wind conditions below the maximum threshold (Cheong et al. 2013).

3.7.1.3 Demographic and socio-economic factors

Population: According to the literature, population density is one of the commonly-used predictors in DF modelling studies. Moreover, both high and low population density is a major factor contributing to dengue epidemics (Cummings et al. 2004; Schmidt et al. 2011). Additionally, increasing urbanization and greater population density provide favourable conditions for the *Aedes* mosquito, which can increase DF transmission (Qi et al. 2015). However, data for other demographic and socio-economic factors including Gross Domestic Product (GDP), housing, education, and water storage methods are unavailable or difficult for the study area.

3.7.1.4 Environmental factors

Elevation: Since elevation and temperature are inversely related, it was predicted that elevation would also significantly affect dengue vector survival based on the widely reported effect of temperature on the species (Ashby et al. 2017). Moreover, vector

breeding is regulated by elevation, which determines other climatic factors (temperature, and humidity) which are directly related to mosquito biting activity, virus incubation, and vector lifespan (Roslan et al. 2016).

Land Use Land Cover (LULC): According to research, land use and land cover (LULC) changes may play a significant role in causing DF (Gao et al. 2021). Moreover, in particular, it was found that the dengue pandemic is positively correlated with vegetation and surface water (Sarfraz et al. 2014; Tian et al. 2016).

3.8 Software used

In this thesis, the main data for dengue patient cases were obtained from the Vector-borne Diseases and Zoonotic of the Health Affairs Department of the Saudi Ministry of Health in an Excel file. The recorded data were in Arabic, which had to be translated into English, and then typo errors had to be fixed before integrating the data into Shapefile. Several software applications were used in the pre-processing stage: ArcMap, QGIS, Excel, and Jupyter notebook. Excel files of DF data were aggregated using pivot tables to extract the total annual number of confirmed cases for each district (2012-2018). Python programming language was used to develop several machine learning approaches to impute the MD as well as to predict DF cases. The cellular automaton (CA) method was applied using the MOLUSCE plugin and QGIS software to simulate potential future threats. Lastly, ArcMap software was used to visualize the simulation risk areas and compare the generated map with the recorded data.

3.9 Summary

The developed DF spatiotemporal prediction models were delineated by the following:

- Jeddah city, Saudi Arabia, was selected as the area of interest on which the DF spatiotemporal prediction model was based.
- This city was chosen for the case study as it has the highest incidence of reported cases of DF in the Kingdom. Moreover, it is the most important port in the Kingdom and is the main gateway for Muslims from all over the world travelling to Mecca to fulfil their religious obligations.
- The main factors contributing to the transmission of DF were identified and collected. These factors include previous dengue cases, climatic "temperature, humidity, precipitation, and wind speed", LULC, and elevation. The collected data were prepared to fulfil the research objectives; develop a data analytical model in the presence of MD, improve the prediction model performance, and perform a simulation of the disease risk areas accurately.
- By means of advanced machine learning approaches, MD were imputed to improve the performance of the DF prediction model and achieve high accuracy prediction under specific circumstances.
- The application of the cellular automata method contributed significantly to the accurate assessment and prediction of DF hotspot areas using district boundaries as the spatial scale and one year as the temporal scale.
- All the algorithms and methods were applied using a GIS-based environment and Python programming language.

CHAPTER 4

RESULTS AND DISCUSSION

4.1 Introduction

This chapter presents and discusses the results from the proposed methodologies used to achieve the thesis objective and improve the DF prediction models from a spatiotemporal perspective. Jeddah city was used as the case study and the district's boundaries were the spatial scale and one year interval was the temporal scale. The results showed the ability of the proposed data analytical model to assess the DF cases and provide a better understanding of related risk factors in the presence of missing data (MD). Moreover, the results illustrate the finding of adopting an advance imputation algorithm to fill in missing values accurately in comparison to traditional imputation approaches. Then, the improvement of DF spatiotemporal prediction models was demonstrated and high accuracy performance was achieved. In addition, the hotspot areas in terms of district spatial scale were predicted along with the major factors influencing the disease transmission. Lastly, a simulation was conducted and high-resolution risk maps were generated for a better understanding of DF transmission in these districts. Figure 4.1 illustrates the flowchart of the thesis's achieved results.

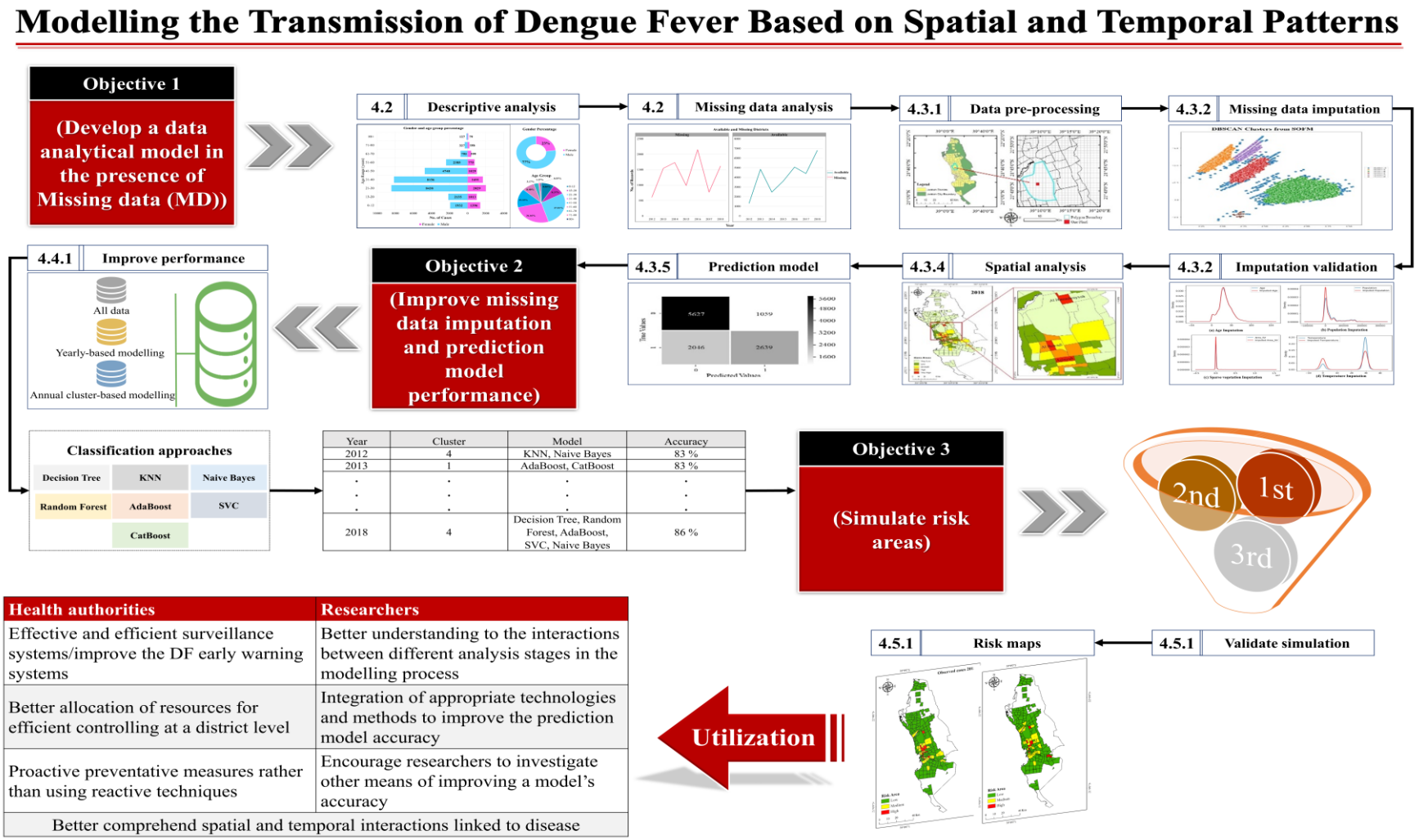


Figure 4.1. Research results flowchart

4.2 Descriptive analysis

A total of 37,903 confirmed/suspected cases of DF were reported to Jeddah Health Affairs from 2012 to 2018. Generally, the number of confirmed cases varied across the years of interest: the largest number of confirmed cases was 4,974 patients in 2018, while the lowest number of confirmed cases (991) was recorded for 2012, as shown in Figure 4.2.

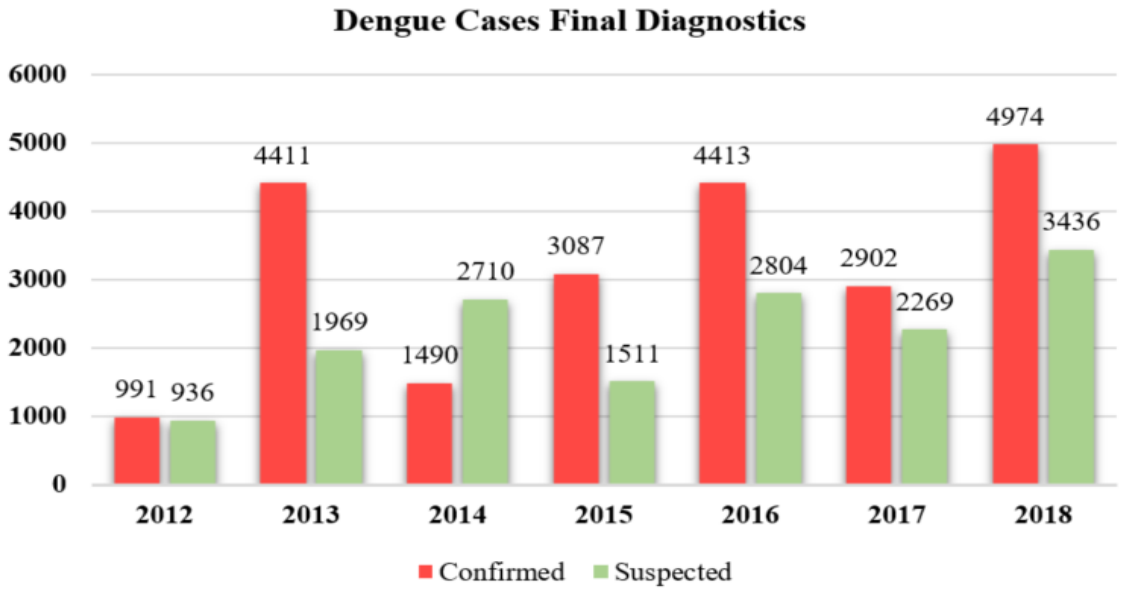


Figure 4.2. Annual reported DF cases

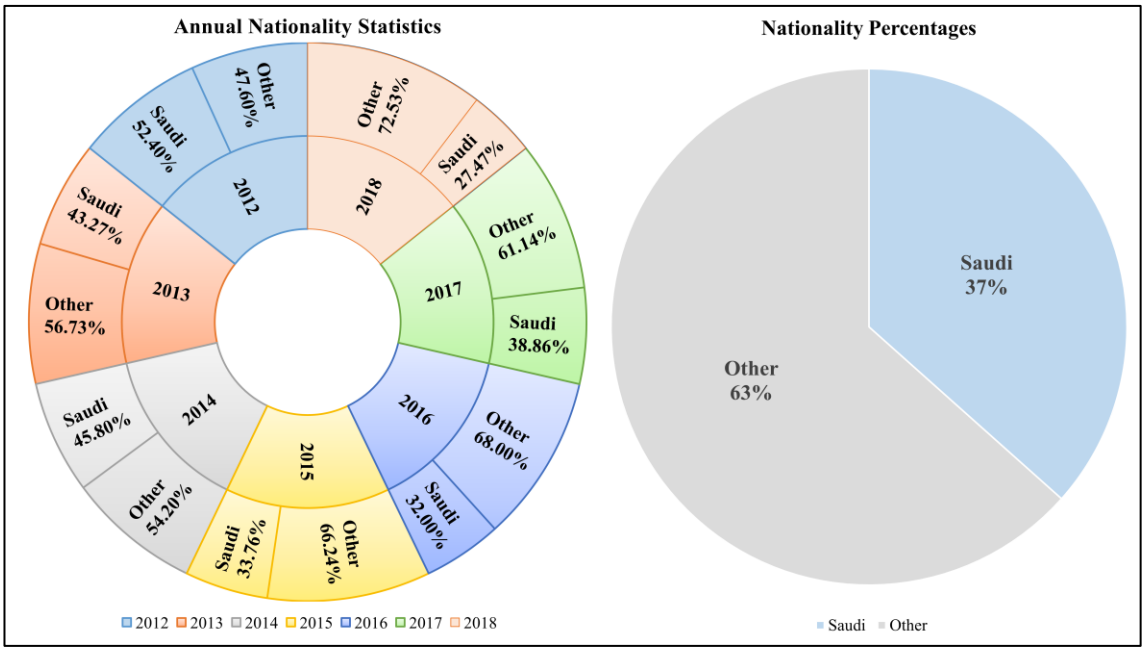
Table 4.1 below is a statistical representation of the results obtained from an analysis of the demographic data. It allows a comparison of the demographic features and their possible impact on disease transmission over seven years. Moreover, it shows fluctuations of confirmed and suspected DF cases at an annual level. Although values were missing for some features, in this study, a reasonable imputation was conducted of these MD to determine the spatial patterns of DF associated with developing a data analytical model in the presence of MD (first objective), and improving the MD

imputation and prediction model performance (second objective), as explained below in Subsections 4.3 and 4.4.

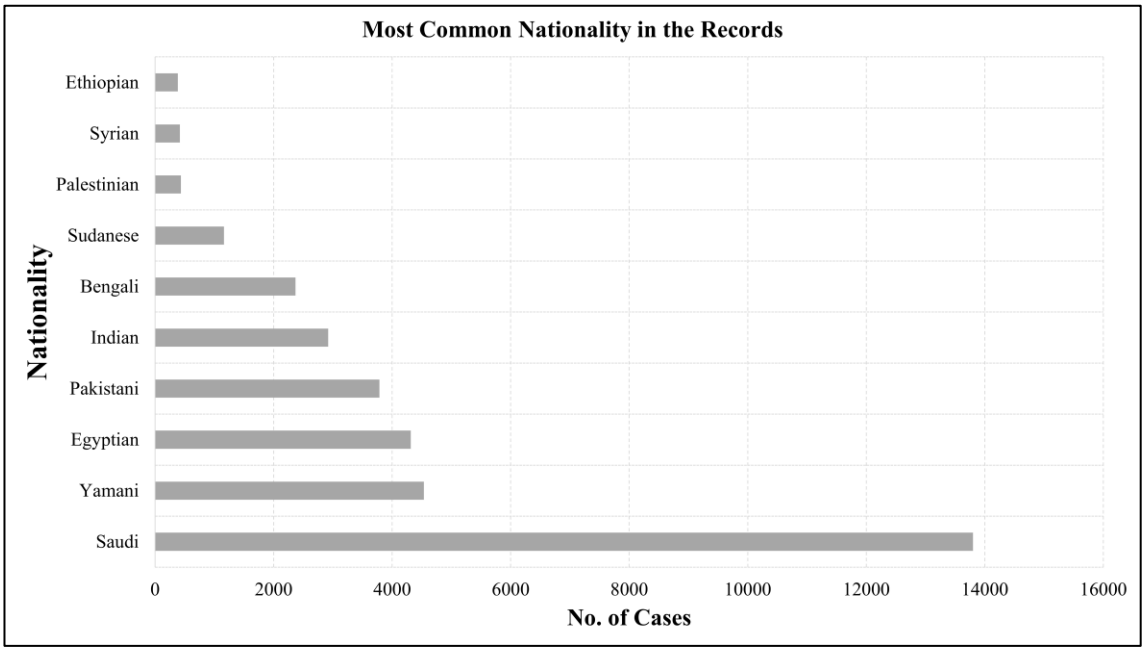
Table 4.1. Descriptive analysis of demographic features

Year	2012	2013	2014	2015	2016	2017	2018	Total
Diagnostics								
Confirmed	991	4411	1490	3087	4413	2902	4974	22268
Suspected	936	1969	2710	1511	2804	2269	3436	15635
Gender								
Female	556	1523	1128	1036	1543	1246	1626	8658
Male	1370	4856	3072	3562	5660	3923	6783	29226
Missing records	1	1	0	0	14	2	1	19
Age								
Average age	35.95	37.01	36.35	38.75	39.79	45.84	46.64	-
Missing records	21	96	48	34	47	1	5	252
Nationality								
Saudi	1003	2731	1909	1544	2303	2004	2307	13801
Not Saudi	911	3581	2259	3029	4012	3153	6089	23034
Missing records	13	68	32	25	902	14	14	1068
District								
Available records	1320	4836	2470	3612	5069	4399	6798	28504
Missing records	607	1544	1730	986	2148	772	1612	9399
City								
Available records	1539	5360	2639	3740	5700	4809	7833	31620
Missing records	388	1020	1561	858	1517	362	577	6283
Total Cases	1927	6380	4200	4598	7217	5171	8410	37903

The demographic analysis shows that the majority of reported cases were non-Saudi (as shown in Figure 4.3 (a)), possibly due to the high number of foreign workers residing in Jeddah city. Most of the non-Saudi reported cases were Yemeni, followed by Egyptians, Pakistanis, Indians and Bangladeshis (Figure 4.3 (b)).



(a)



(b)

Figure 4.3. (a): Nationality statistics, and (b) nationalities with the highest number of recorded infections

The analysis of gender and age groups revealed that most of the recorded cases (more than 77%) were male, as shown in Figure 4.4. This might be due to Saudi culture requiring women to be fully covered outdoors, and men are more likely than women to

work outdoors (Badreddine et al. 2017). All age groups of this study were affected by dengue, but most of the cases occurred in the 21–40 age group (Figure 4.4).

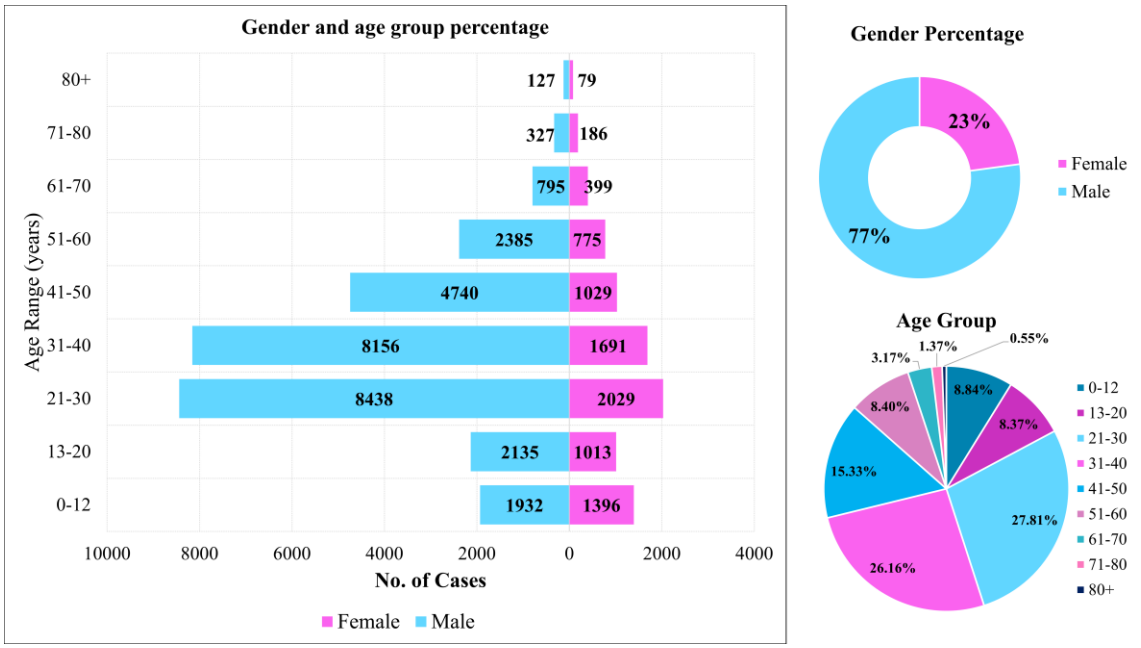


Figure 4.4. Gender and age percentages

Factors other than those related to demographics, such as climate, elevation, land use and land cover are available for Jeddah for the period being investigated, and can be imputed by matching these factors to the exact district "polygon" by means of ArcMap 10.4. However, of the 37,903 reported cases, 1,423 were visitors from other cities, and there is no indication of whether they became infected in Jeddah or in their local region, such as Mecca and Medina, where the epidemic might be more widespread. The areas with the highest number of reported cases are Mecca (547 cases) and Rabigh (526 cases) respectively. To determine the impact of mass gatherings on the transmission of the disease, a new column was added, "Hajj Days", to show the number of cases reported during these days. Of the 98 reported cases for this period, 51 were confirmed. Hence, 98

cases out of 37,903 is a relatively small number and does not suggest any significant link between *Al-Hajj* and the incidence of disease.

The district name is the most important feature as it enables the risk areas to be identified, together with the associated risk factors. As shown in Figure 4.5, annual statistics show that there are always cases for which district information is missing. As explained later, various methods were applied to fill in the missing values.

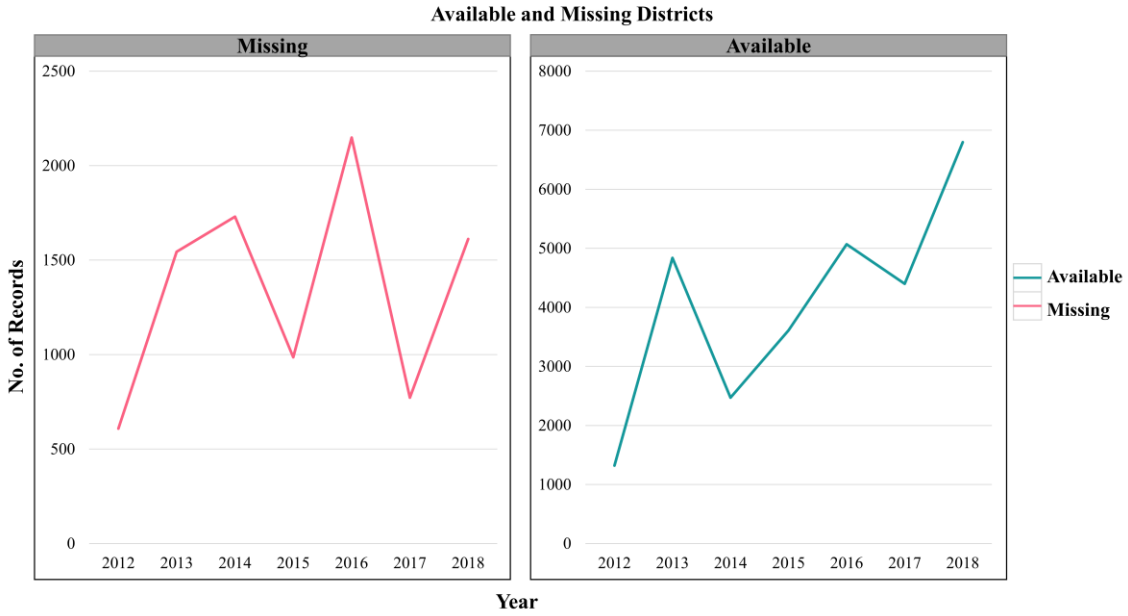


Figure 4.5. Annual records statistics based on missing and available district names

4.3 Objective 1 (Develop a data analytical model in the presence of MD)

4.3.1 Data pre-processing

The patients' data were recorded in the Arabic language, and contained numerous errors due to abbreviations, incorrect spelling, typos, districts named with old common names not officially documented and, finally, information being entered in the wrong column, such as the nationality being placed under occupation. Since the names of some other city districts' are similar to Jeddah districts, new columns were created to avoid the

overlap of districts' names (e.g., "municipality" and "city") outside Jeddah when recording the address. Unfortunately, patient data does not contain a postal code and is not associated with the national address, which may hamper the process of determining the appropriate neighbourhood for each case. Therefore, all errors relating to districts and addresses were corrected manually via Google Maps. However, after the manual input of district names, more than 24.7% of the reported cases were missing location data because the address had been misreported. Hence, pre-processing is essential to check the data, unify values, and improve the data analysis outcomes. Table 4.1 lists the demographic items (age, gender, nationality, and district) with the highest number of missing values in the dataset. The data contained discrete information such as age and recording year, and information on fluctuating factors such as population, temperature, humidity, wind speed, precipitation, and other environmental features. Therefore, normalization was conducted at this stage to unify the range of these features by transforming the values in numerical columns by scaling each to a given range between 0 and 1 by applying "MinMaxScaler", adopted from sklearn. Since the data contained categorical variables, two categorical encoding approaches were performed. The dependent variable “The Final Diagnosis" contains two unique values: "Confirmed" and "Suspected". So, label encoding was conducted for this feature by allocating (1) to confirmed cases and (0) to suspected cases. Four independent features (nationality, municipality, city, gender, and district) were encoded by means of binary encoding since some of the independent features (nationality, district) contained a high number of value categories. After the normalization and encoding processes, all nominal variables were handled as numeric variables in the range of 0 to 1.

Spatial and temporal scales criteria

The collected data on notified cases were observed in an individual format, including the recorded date and the district and municipality names for each case. The district name was considered as the finer spatial scale for analysis, as the municipality encompasses multiple districts, making it a more general representation of hot spots. The temporal scale chosen for the analysis was one "year," as the climatic and land use land cover (LULC) data are slightly changing over time. Thus, in order to fulfil the research objectives and to gain a deeper understanding of the spatial-temporal characteristics of the disease, the analysis was carried out at the district level in Jeddah City as a spatial scale, and a period of one year was selected as the temporal scale. However, as discussed in Section 5.7, it is recommended that future studies consider different spatial and temporal scales based on the specific data sets used.

4.3.2 MD imputation

MD is a major problem that reduces data quality and the performance of a model. For this objective, an imputation method was adopted and applied to address this issue and ascertain the most appropriate model for the dataset. Therefore, following the pre-processing stage, the output data was used to train the Self-Organizing Feature Map (SOFM or SOM). SOFM was adopted due to its ability to simplify complex, high-dimensional data and reduce its dimensions without affecting the relationship among data variables (Asan and Ercan 2012). Thus, two dimension methods were used to visualize and investigate the relationships in the dataset. All the values of the factors in the dataset, including missing values, were encoded, as explained in the data pre-processing stage, to train the SOFM. To train the SOFM, a loop function containing default values in addition to the values of the main parameters such as learning radius was implanted, “std” which

is a particular parameter used to control the learning rate of each neighbour, the features grid, grid type, and a learning rate "step". This process was conducted separately for each year from 2012 to 2018 to avoid the overlapping of the data visualization and to determine the relationship among data variables on a yearly basis.

The DBSCAN processes the mapping data obtained by SOFM to analyse the data clusters. As explained in the methodology section, this approach is easy to implement since only two parameters are considered, the distance between two points to include in the cluster "eps" and the minimum point at each cluster "$MinPts$". Table 4.2 shows the best parameters for SOFM and DBSCAN based on a specific year, using the hyperparameter method. Figure 4.6 shows the clusters obtained by SOFM-DBSCAN.

Table 4.2. Parameter values used for SOFM and DBSCAN

Year	SOFM					DBSCAN	
	Reduce radius After	Reduce Std After	Reduce Step After	std	step	eps	MinPts
2012	100	1	1	1	1.09	0.095	10
2013	100	1	1	1	1.1	0.095	23
2014	100	1	1	1	1.09	0.095	5
2015	100	1	10	1	1	0.095	11
2016	100	1	1	1	1	0.095	9
2017	100	1	1	1	1	0.095	40
2018	100	1	1	1	1	0.095	6
All Years *	100	1	10	1	1	0.05	10
Complete Data **	100	1	50	1	1	0.05	25

* All Years refers to all data used together in the analysis. ** Complete data represents the data without missing values by ignoring these values

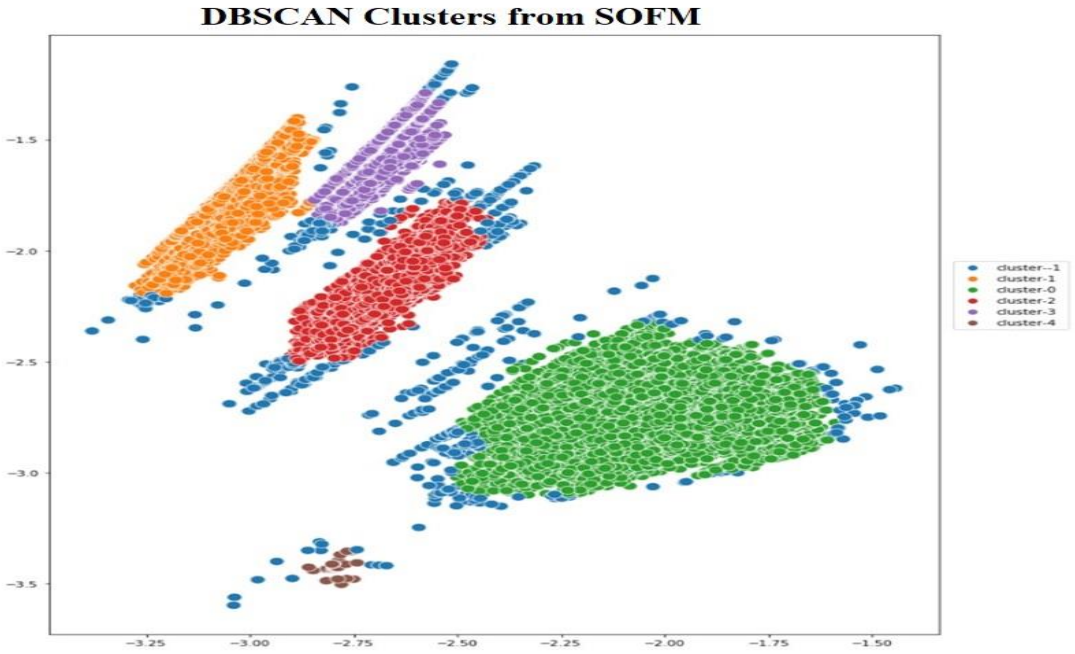


Figure 4.6. DBSCAN clusters obtained from SOFM for all annual datasets

The imputing of MD was done differently for numerical and categorical features. However, this process was conducted for each cluster extracted by SOFM and DBSCAN as explained in the following section. Several methods were applied to fill in MD for features with numerical values such as age, population, temperature, and humidity. These imputation methods were: mean, median, and mode values. Following these four methods, a variance table was created to determine the best way to fill missing values for that feature. All features, except for demographic ones, can be matched again using the collected data to ensure the accuracy of missing values in the variance check. Moreover, a variance check was applied at this stage to compare the variance of imputed values with the original variance. Figure 4.7 (a-d) shows the variance input of various factors. Imputation is better with the lowest variance. Therefore, based on the variance values of used methods, imputation of the mean as the numerical value was the method applied due to the lower values obtained by this method compared to other approaches.

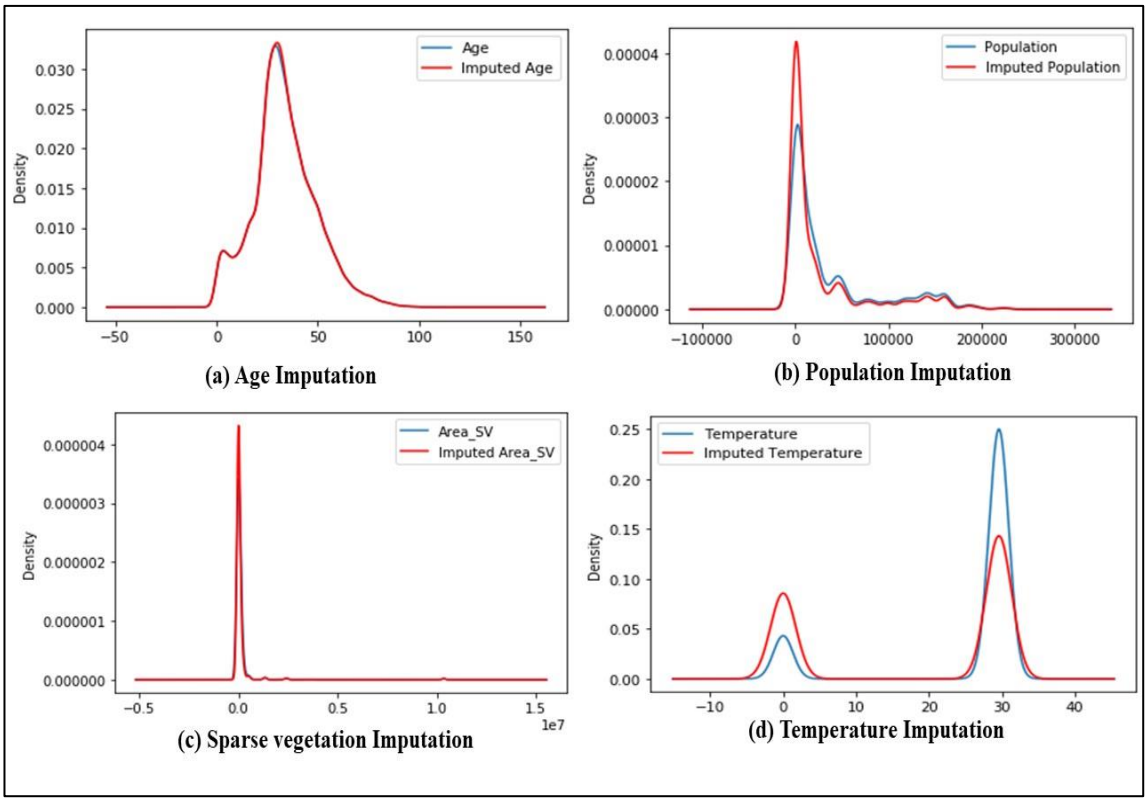


Figure 4.7. Variance comparison of existing and imputed values

For features with categorical values, binary encoding and one-hot encoding were applied and compared to transform these values into numerical values. By doing so, these parameters are taken as numerical attributes and treated accordingly. Prior to encoding, several features were minimized to produce lower values. For instance, the nationality parameter was reduced to Saudi and Non-Saudi instead of stipulating exact nationality. Then, based on the cluster obtained by SOFM and DBSCAN, each missing value in that cluster was filled with the most frequently-occurring value in that cluster. However, a few patient records contained many missing values for most of the features; these records were excluded from the dataset. After imputing the MD, a prediction model was applied to each cluster individually as explained in Section 4.3.5.

4.3.3 Results of Ordinary Least Square (OLS) and Geographically Weighted Regression (GWR)

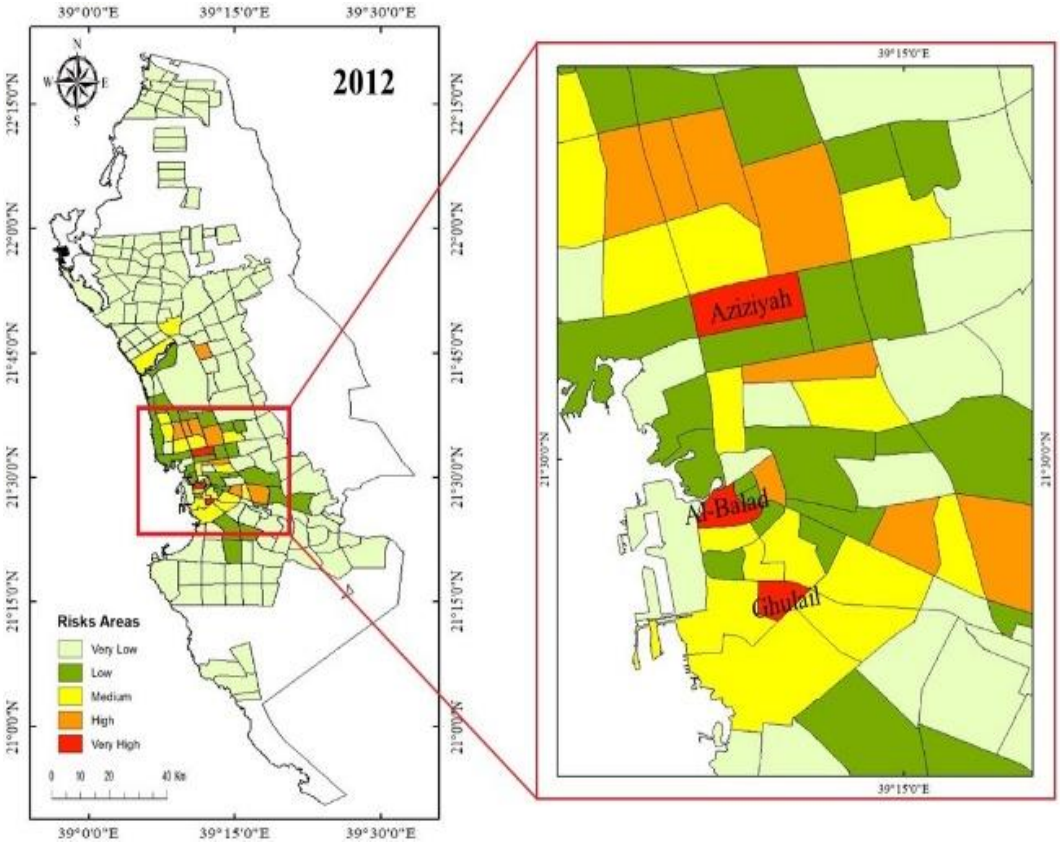
In these two models, the confirmed cases of DF were the dependent parameter and other factors were assigned as independent "explanatory" parameters. In order to tackle multicollinearity issues, only those variables with a variance factor inflation of less than 7.5 (VIF < 7.5) were included in the model. Gender, nationality, and temperature (except for temperature in 2012) were the variables that exhibited multicollinearity. Therefore, the OLS model was applied several times to the annual statistics to determine the significant parameters fitting the model. Following this, the GWR was performed for each year using the same dependent and independent parameters as those for OLS. However, in this approach, multicollinearity variables are removed, since it is more sensitive to these kinds of parameters and produces errors. Both approaches were evaluated to determine the best-fit model using several measurement methods: Akaike Information Criterion (AIC), Adjusted R^2, and R^2 as shown in Table 4.3. The results of the three measurements indicate that both methods produce almost similar values. However, after comparing the results obtained by the two approaches, OLS is more appropriate for the collected data as the values of AIC are smaller than the values of AIC in GWR.

Table 4.3. Comparison of OLS and GWR results

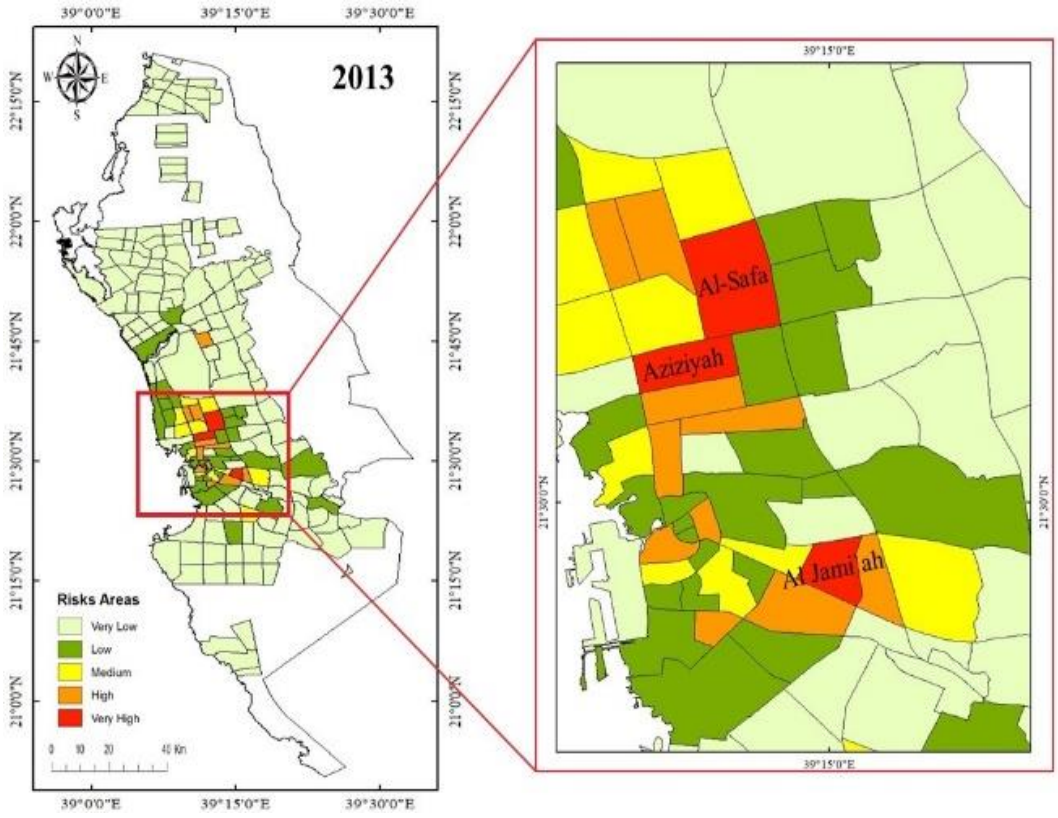
Year	OLS			GWR		
	AIC	Adjusted R^2	R^2	AIC	Adjusted R^2	R^2
2012	1329.2503	0.3402	0.3887	1329.2602	0.3401	0.3887
2013	1946.6543	0.3808	0.4264	1946.6640	0.3808	0.4265
2014	1457.3274	0.3388	0.3874	1457.3402	0.3388	0.3874
2015	1903.5940	0.2741	0.3275	1903.6034	0.2741	0.3275
2016	2009.0245	0.2415	0.2973	2009.0389	0.2415	0.2973
2017	1863.6095	0.2350	0.2912	1863.6230	0.2350	0.2913
2018	2091.7213	0.2449	0.3004	2091.7364	0.2449	0.3005

4.3.4 Spatiotemporal analysis of DF risk areas from 2012 to 2018

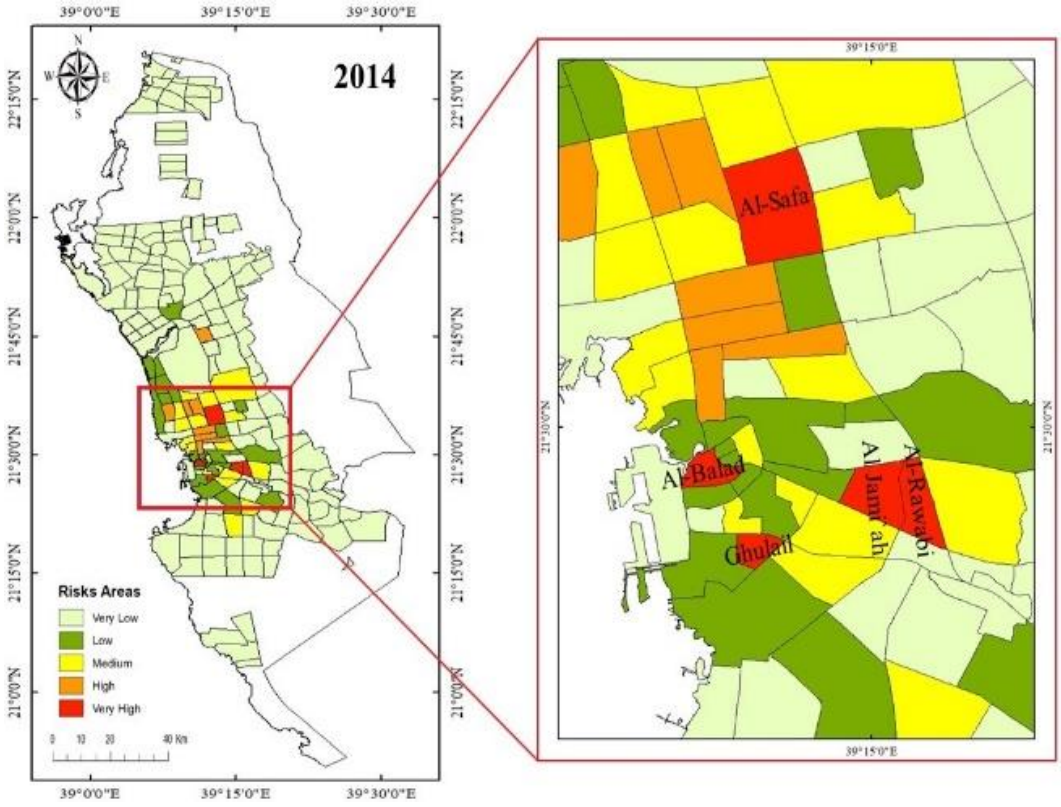
The data for reported cases of DF from 2012 to 2018 indicate that, given the number of confirmed cases, the central districts in Jeddah are the hotspot areas. The red-coloured areas in Figure 4.8 indicate the annual hotspots (temporal scale) based on confirmed cases; orange shows the second-highest risk areas. The green indicates the "cold spot" districts where there were fewer reported cases. Although, in this study, the high-risk areas are mainly in the central districts of Jeddah city, the number of confirmed cases fluctuates yearly from one district to another. For example, Ghulail had the highest incidence of confirmed cases in 2012 and 2014, while the Al-Safa district had the highest number in 2013 and 2015. The Al Hamadaniyyah district had the highest number of confirmed cases for the period spanning 2016 to 2018.



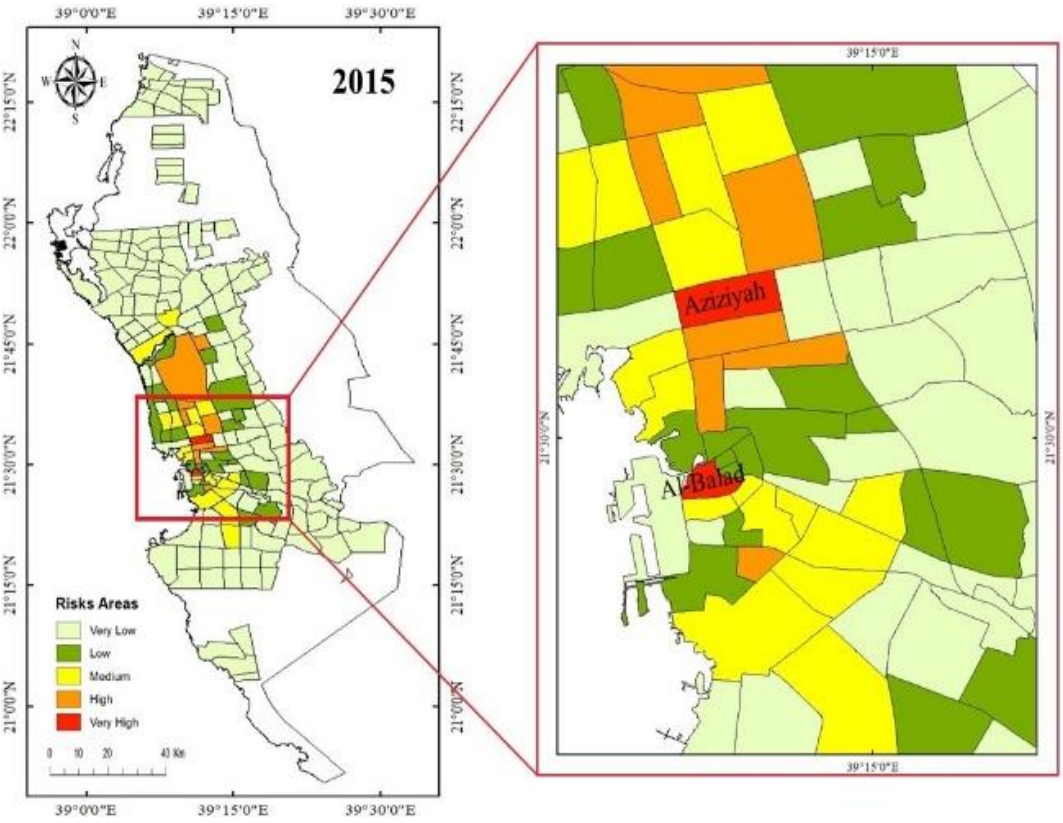
(a) The year 2012 high-risk districts with notified confirmed DF cases



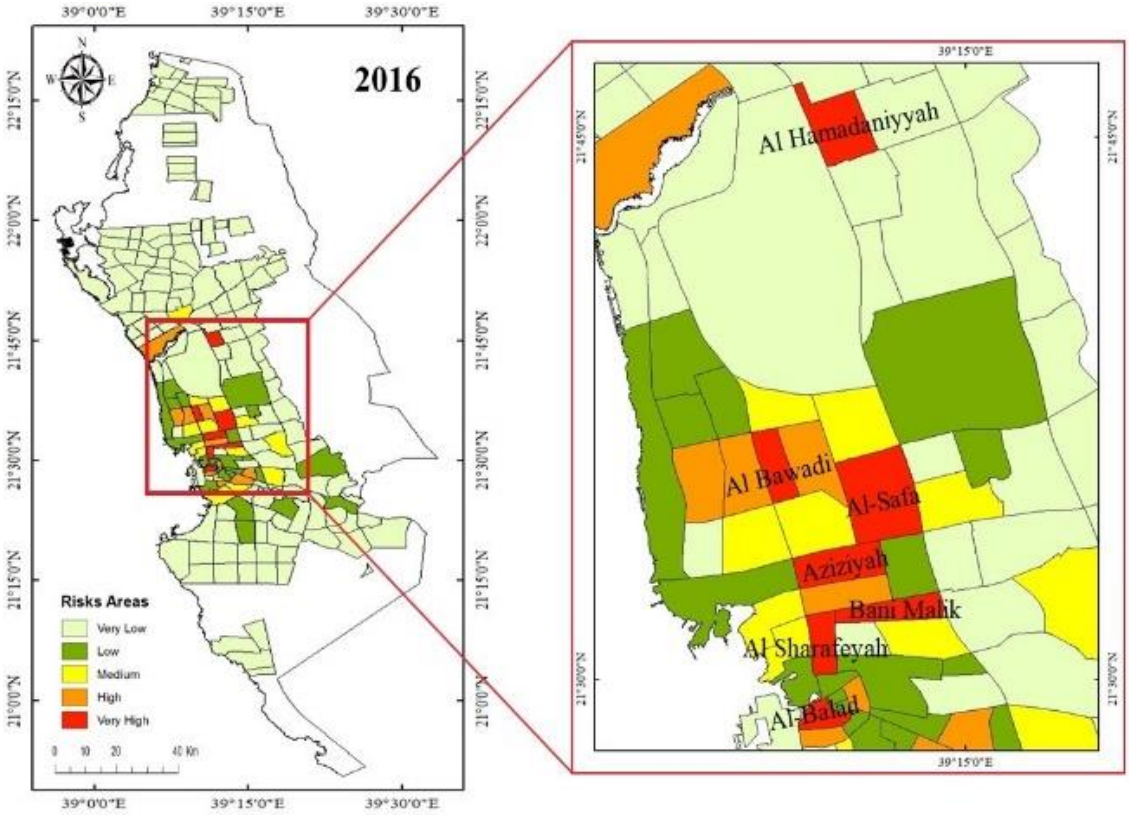
(b) The year 2013 high-risk districts with notified confirmed DF cases



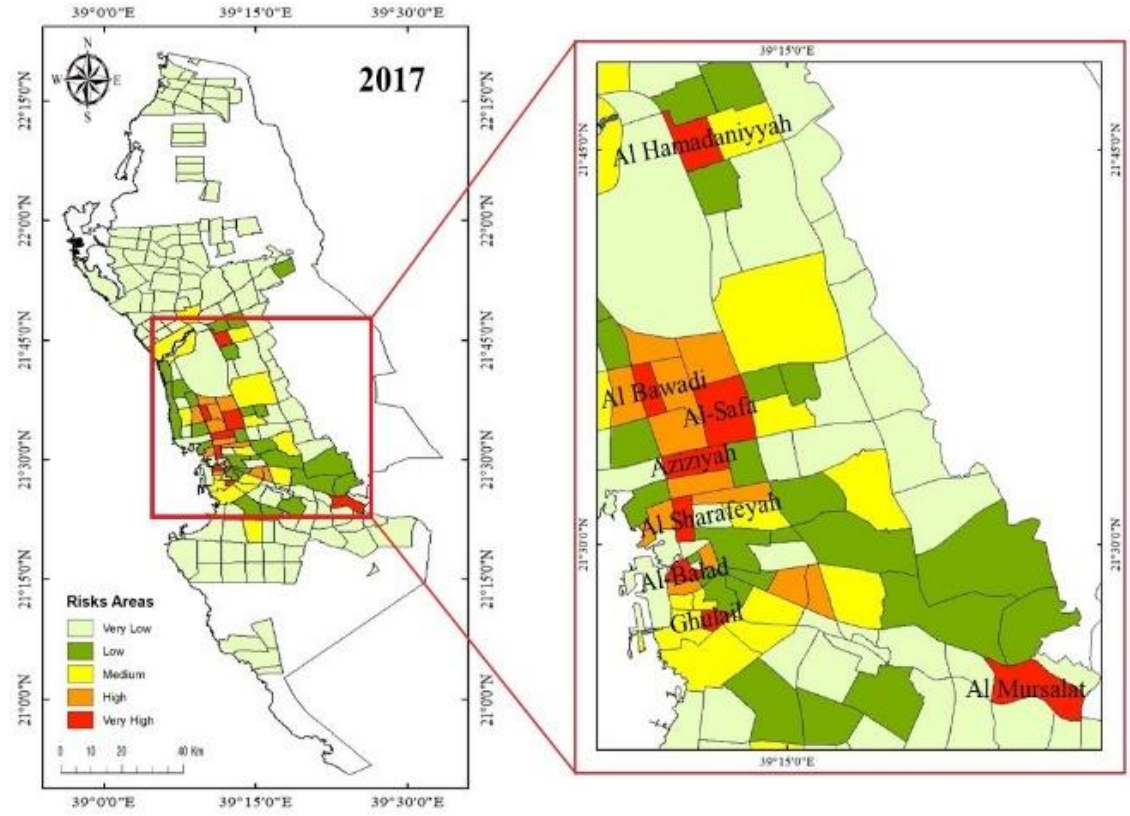
(c) The year 2014 high-risk districts with notified confirmed DF cases



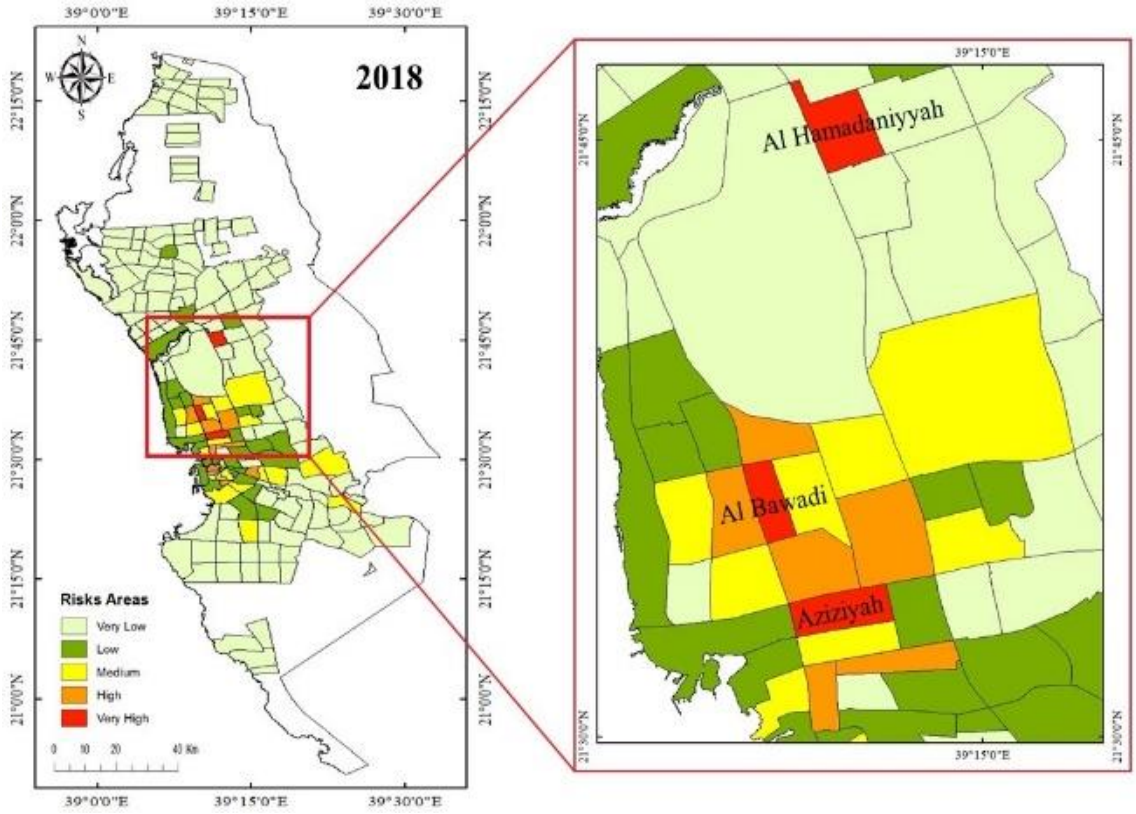
(d) The year 2015 high-risk districts with notified confirmed DF cases



(e) The year 2016 high-risk districts with notified confirmed DF cases



(f) The year 2017 high-risk districts with notified confirmed DF cases



(g) The year 2018 high-risk districts with notified confirmed DF cases

Figure 4.8. High-risk districts with notified confirmed DF cases

Although the most dangerous (i.e., hotspot) areas vary from year to year, they are usually one of the high-risk areas in any one year in terms of confirmed DF cases. The hotspot areas with high numbers of confirmed cases were identified using ArcMap 10.4 software (Getis-Ord Gi*), and the danger zones are in the middle of the study area of interest, as shown in Figure 4.9.

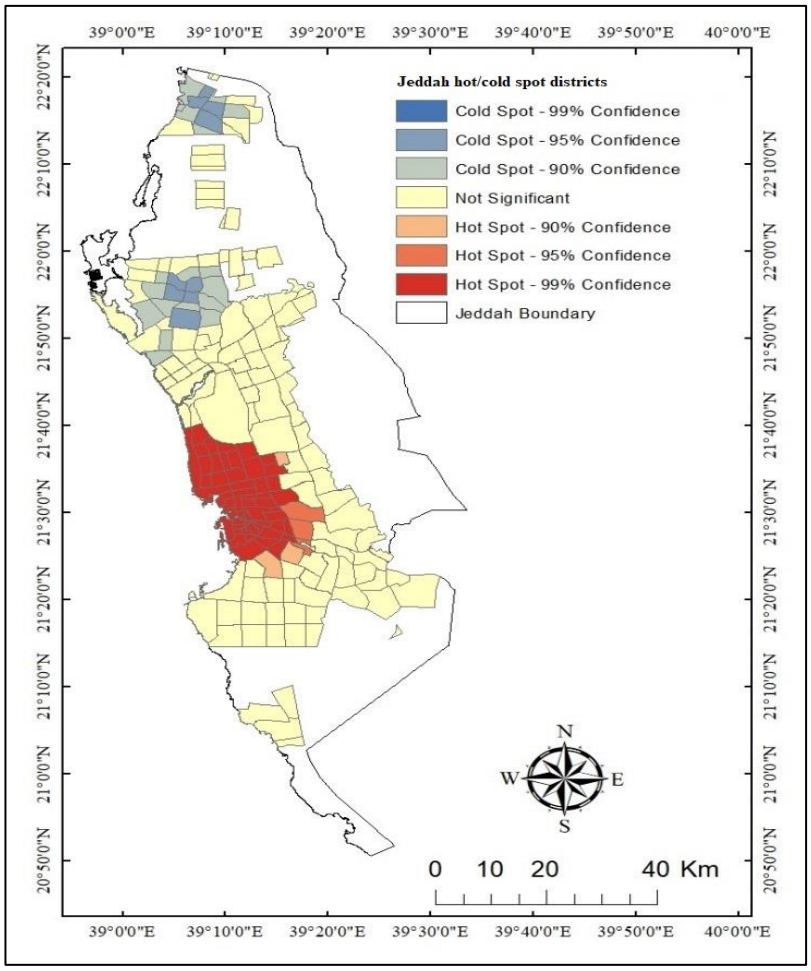


Figure 4.1. The hot/cold spots (Getis-Ord Gi*) results for 2012

The temporal risk of the epidemic was indicated by the annual figures for confirmed cases and the months with the highest numbers of DF cases. Table 4.4 gives the number of confirmed cases monthly and yearly, along with the total number of cases on that scale. It was noted that the number of confirmed cases increased more than three times in 2013 compared to 2012, and then decreased in 2014 to reach 1490 confirmed cases, and then began to increase again to 3075 and 4413 confirmed cases in 2015 and 2016, respectively. Finally, the number of confirmed cases in 2017 decreased from the previous year to 2846 cases, before rising again to 4952 cases in 2018, recording the largest number of confirmed cases for the study period as shown in Figure 4.10.

Table 4.4. Monthly/Yearly confirmed cases

Month	2012	2013	2014	2015	2016	2017	2018	Total
January	52	304	80	131	159	233	213	**1172**
February	26	271	91	157	134	212	218	**1109**
March	30	441	120	216	289	206	405	**1707**
April	50	809	159	376	615	297	1100	**3406**
May	133	1076	258	493	987	555	1183	**4685**
June	209	793	297	746	935	576	989	**4545**
July	166	358	189	390	575	298	465	**2441**
August	66	100	73	206	219	122	114	**900**
September	30	44	38	62	135	50	109	**468**
October	28	38	44	49	94	61	46	**360**
November	64	50	64	51	98	68	42	**437**
December	137	127	77	198	173	168	68	**948**
Total	**991**	**4411**	**1490**	**3075**	**4413**	**2846**	**4952**	**22178**

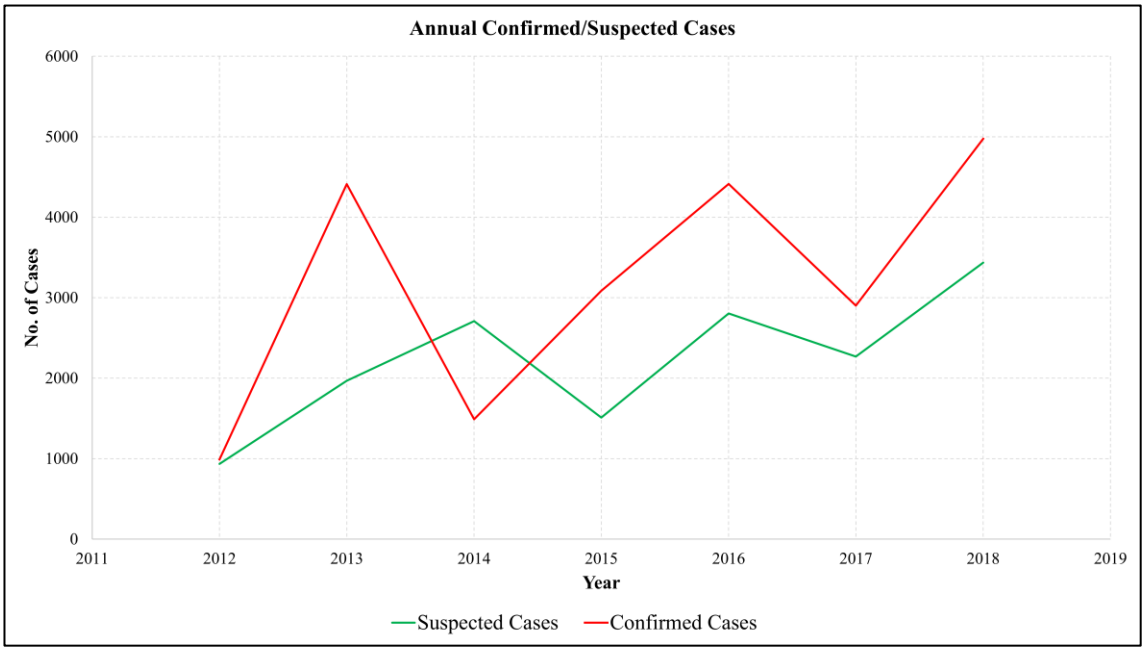


Figure 4.10. Confirmed DF cases annually

Figure 4.11 shows, beginning from March, the number of confirmed cases increases until May when it peaks. Then the number decreases slightly in June, before it decreases significantly until October, when the lowest number of cases are recorded for the year.

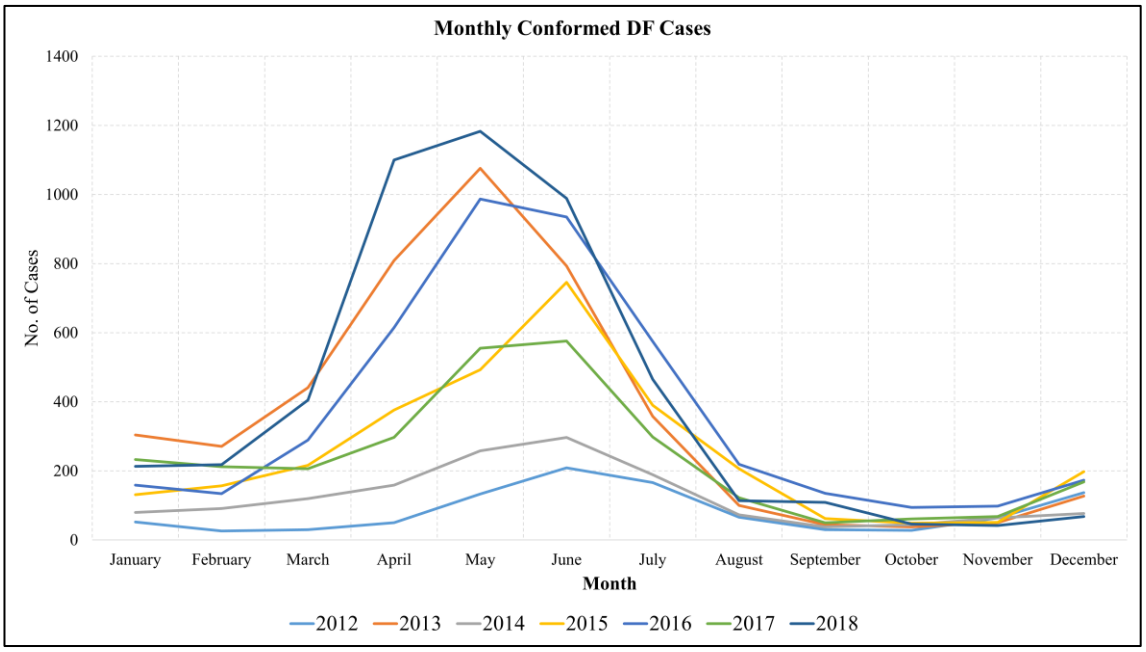


Figure 4.11. DF confirmed cases monthly

4.3.5 Prediction of confirmed DF cases using CatBoost classifier

According to Eq. 1, the CatBoost model achieved an accuracy of 73% for the prediction of DF cases as illustrates in Figure 4.12. Compared with other common machine learning (ML) approaches, the CatBoost model yields the most accurate prediction results. The models applied to each cluster individually as well as to all clusters simultaneously, are shown in Table 4.5. The impact of data size, based on the created clusters using "SOFM and DBSCAN" was compared to determine the performance of the various models of interest. Lastly, several ML approaches were adopted and results were compared to identify the model that performed best on the data. These approaches were: Decision Tree, KNN, Random Forest, AdaBoost, Support Vector Classification (SVC), and Naive Bayes. Based on comparing the models' performances when they were applied to all clusters at once: KNN Classifier, Random Forest, and AdaBoost achieved 66% accuracy, while Decision Tree, SVC, and Naive Bayes achieved 64%, 59%, and 57% respectively. However, the models' performances improved when applied to each cluster

individually instead of considering all of the clusters at once; for instance, KNN and random forest achieved 74% for cluster 4, while 66% accuracy was obtained when applied to all clusters at once. Hence, further investigation was conducted in the second objective as discussed in Subsection 4.4.

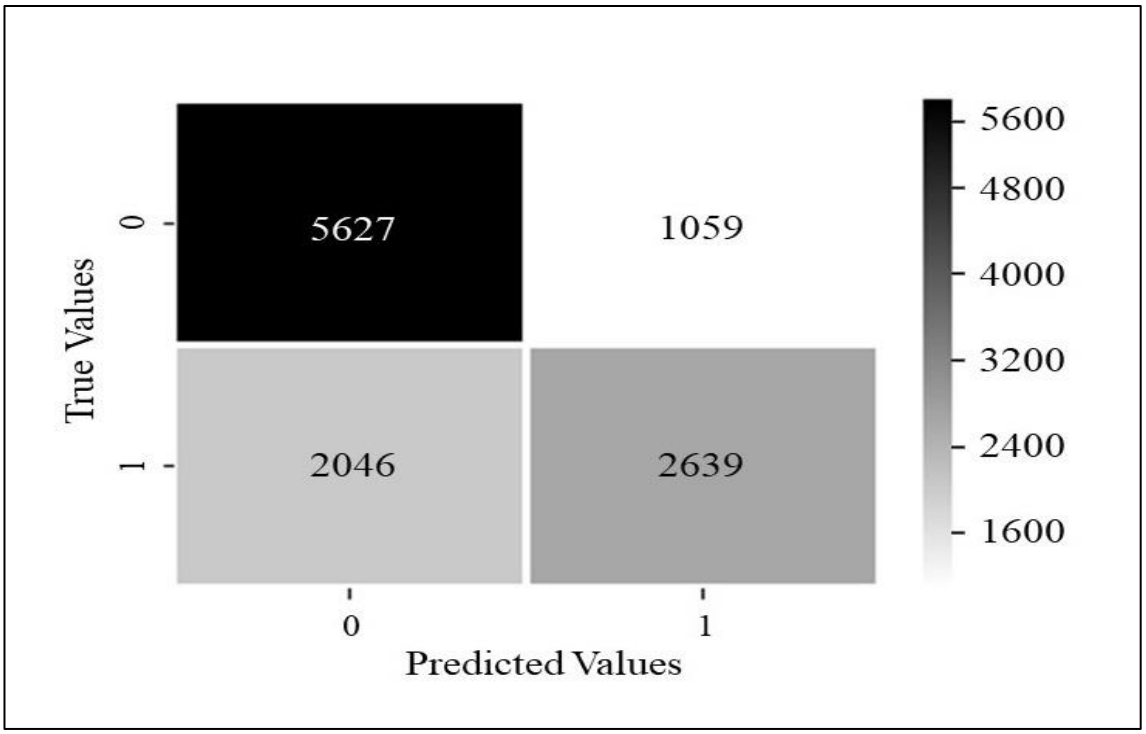


Figure 4.12. Catboost model confusion matrix

Table 4.5. Comparison of model accuracy

Cluster	Cluster Size	Model						
		Decision Tree	KNN	Random Forest	AdaBoost	SVC	CatBoost	Naive Bayes
-1	614	0.66	0.58	0.63	0.61	0.5	0.68	0.5
0	23610	0.63	0.66	0.64	0.67	0.6	0.73	0.58
1	7677	0.61	0.61	0.61	0.63	0.59	0.69	0.59
2	4374	0.65	0.66	0.67	0.71	0.67	0.72	0.54
3	1514	0.58	0.62	0.63	0.68	0.66	0.68	0.6
4	114	0.71	0.74	0.74	0.69	0.69	0.71	0.69
All	37903	0.64	0.66	0.66	0.66	0.59	0.73	0.57

4.3.6 Discussion

This study utilized GIS and satellite images to integrate DF cases, and climatic and non-climatic risk factors for spatiotemporal analysis, in order to help address DF in Jeddah city, which remains a significant public health concern. Overall, the results highlight the risk of DF in terms of both spatial and temporal scales, and the proposed model predicts the potential causes of the disease. To the best of our knowledge, this is the first study to adopt ML methods to predict DF cases in 205 districts in Jeddah.

A spatial analysis covering the study period found that the disease hotspots were located in the centre of the investigated area, similar to the findings of previous studies (Khormi et al. 2011). Although the high-risk areas are mainly in central districts, the monthly and annual data show that, based on the highest number of confirmed cases, the danger zones change over time and are not restricted to a particular district, which is identical to the results obtained by a previous study of the same area (Khormi et al. 2011), and another study conducted in Australia (Naish and Tong 2014). In this current study, regardless of whether raw DF counts or DF densities were employed in the calculations, the positions and distributions of hot and cold regions change substantially from year to year. The analysis shows that the Ghulail district was the highest risk zone in 2012 and 2014 based on the confirmed reported cases, while the Al-Safa district was the highest risk zone in 2013 and 2015 and, lastly, the Al Hamadaniyyah district was the highest-risk hotspot area from 2016 to 2018. Demographic data showed that non-Saudi people accounted for the highest number of confirmed cases, possibly because they constitute the highest proportion of the population in the middle- and high-risk areas (Khormi and Kumar 2011). Similarly, a demographic analysis of the disease in the State of Queensland, Australia, showed that the average age of reported cases was 38 years. Moreover, in terms of gender, there was no significant difference since 49.7% of the total reported cases were

male (Viennet et al. 2014). The results in regard to gender are significantly different since, in the Jeddah district, 77% of reported cases were male. This may be due to the Saudi culture where women must be covered outdoors, and where men are more likely to have jobs requiring them to be outdoors (Badreddine et al. 2017). As Table 4.1 shows, apart from 2014, every other year for the study period shows an increase in the average age of reported cases, from mid-thirties in 2012 to mid-forties in 2018.

The significant factors that contribute to the transmission of the disease, including climatic and environmental factors, were derived from previous studies. However, prior to applying spatial analysis using OLS and GWR, the CatBoost open-source package was used to extract the significant factors for the proposed prediction model. Moreover, determining the important features will decrease the training time and retain the most significant parameters for the analysis (Al-Sarem et al. 2021). The significant factors were used as explanatory parameters in the OLS and GWR models, while the DF cases were the dependent parameter. Based on the observed data, OLS was more appropriate for the data and was applied for the spatial analysis, unlike previous studies where GWR was found to be more suitable (Acharya et al. 2018; Khormi and Kumar 2011).

Although MD is one of the most common issues directly affecting data quality, it is sometimes handled inappropriately by ML methods, making that data useless (Batista and Monard 2003). To reduce bias in the data, this problem must be dealt with carefully, taking into account the mechanism most appropriate for the data (Batista and Monard 2003). During the pre-processing stage, after translating the observed data from Arabic to English, thousands of records were identified as unclear or missing the address or district. Hence, this information was checked manually through Google Maps and the MD was filled in as in a previous study (Viennet et al. 2014). Gaps in the observed data can be filled by matching these records with open-source data (Buczak et al. 2014).

Unlike previous studies, this study adopted a different approach to fill in the MD; the "mode" value in each cluster was found by SOFM and DBSCAN methods to reduce any bias in the data; the mean was used for numerical features as it provides the lowest variance compared with others methods, or these records were excluded under particular circumstances. A common approach, especially when there is a small number of records, is to ignore some of the records in order to achieve the desired goals, particularly if this will not affect the performance of the model (Fuentes-Vallejo 2017; Wiese et al. 2019).

To reduce the future harm caused by the dengue virus, it is necessary to deliver timely predictions of dengue disease. The records maintained by clinics provide valuable data about infected patients. By retaining these extensive data, the possibility of future infection of the general population before transmission occurs can be anticipated (Nithyaa et al. 2019). Most of the previous studies were conducted using a secondary dataset from various demographic regions. Furthermore, as far as researchers' know, no recent research on DF prediction has been conducted in Jeddah city using ML approaches. Sarma (Sarma et al. 2020) applied decision tree and random forest to predict the epidemic using different attributes for 209 records that contained the medical history of patients, and clinical and demographic data. The decision tree achieved better accuracy than the random forest with 79% prediction accuracy. Another study that used SOM to predict dengue and determine model performance achieved only 70% accuracy (Faisal et al. 2010). An appropriate means of investigating a model's performance is to apply and compare various algorithms used for the same datasets (Onan 2019). Therefore, in this current study, several methods were applied to predict the incidence of DF patients from the collected data; they are: CatBoost, Decision Tree, and Naive Bayes. By using CatBoost, the proposed approach obtained better performance than other methods, achieving 73% accuracy. Hence, an effective classifier approach is essential to ascertain

the relationship between factors (Onan et al. 2016). The approaches adopted here to impute MD and perform modelling on each cluster generated by SOFM and DBSCAN, have been applied previously to different datasets and diseases (Shukla et al. 2018). Moreover, these researchers found that grouping the patients into different clusters and modelling each cluster individually improves the prediction accuracy. In this study, the proposed model achieved 68% - 73% accuracy when modelling the clusters separately, while (Shukla et al. 2018) achieved results ranging between 62.82% and 86.84%.

The findings from this study show the risk zones of the epidemic in Jeddah city as well as their relationship with the observed data. Because there is no known treatment for this disease, it is essential that it be contained. Although the proposed model does not explain the spatial-temporal transmission patterns of the disease based on the observed data, it provides significant maps with hotspot areas on an annual temporal scale, which is valuable to health authorities in their efforts to eliminate the disease. At various stages of analysis, appropriate methods were used for filling in missing values to ensure non-bias in the data and to maintain data quality. With the proposed framework, the prediction of DF cases can be achieved with reasonable accuracy.

4.4 Objective 2 (Improve MD imputation and prediction model performance)

4.4.1 Results of MD imputation methods

Approaches based on both statistical and ML methodologies were used to impute missing values in the data of DF patients. The purpose was to examine how the various techniques for imputing MD values improved modelling accuracy. ML models were used to predict confirmed DF cases, and various imputation strategies were examined to determine whether and/or how these strategies influenced the prediction. The methods described in the methodology section were used to fill in any missing values, enabling

comparisons to be made of the various techniques used to impute data. The imputation techniques were applied to local dengue data provided by the Saudi Ministry of Health for the period between 2012 and 2018. Table 4.6 shows the number of records for each year.

Table 4.6. Yearly recorded cases

Data	Records	Complete records*	Records with missing values
2012	1927	1267	660
2013	6380	0	6380
2014	4200	2405	1795
2015	4598	3255	1341
2016	7217	4608	2609
2017	5171	4069	1102
2018	8410	6579	1831
Total	37903	22183	15720

* Complete records refer to data without missing values in all features

To evaluate the performance of the adopted models using different imputation approaches, three scenarios were examined and analysed:

1) All seven years (2012-2018) together in one file
2) Yearly-based modelling
3) Annual cluster-based modelling

Two files were created for the first scenario. One contained complete data obtained only by ignoring the missing values (21,183 records out of 37,903). The second file contained all data, including records with missing values. Table 4.7 shows the performance of the desired models when all seven years were considered together. Here, it is clear that the accuracy of all models is somewhat similar using different imputation strategies, although the CatBoost model performed best, achieving an accuracy of 75%.

Table 4.7. Model performance using all data records

Model	Model performance				
	Complete records *	Mean	Median	Mode	SOFM and DBSCAN
Decision Tree	66 %	65 %	66 %	66 %	64 %
KNN	58 %	61 %	61 %	61 %	65 %
Random Forest	72 %	74 %	74 %	73 %	67 %
AdaBoost	68 %	69 %	69 %	69 %	66 %
SVC	58 %	61 %	61 %	60 %	59 %
CatBoost	75 %	75 %	75 %	75 %	72 %
Naive Bayes	60 %	61 %	61 %	60 %	57 %

* Complete records refer to keeping complete data only and discarding all records with missing values.

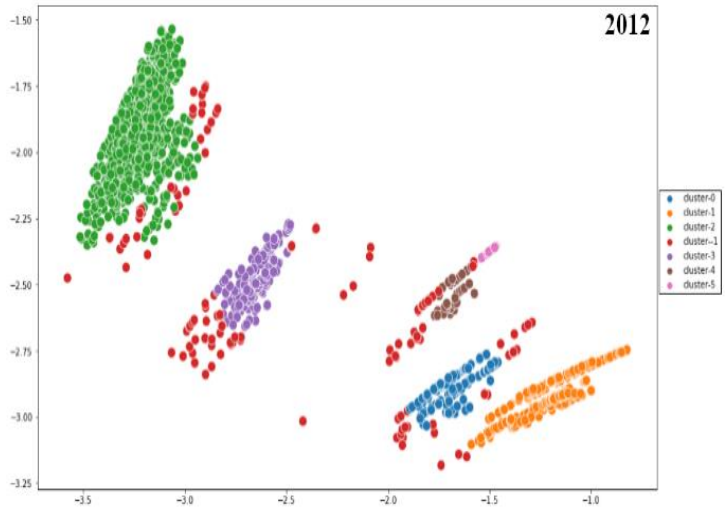
For the second scenario (Table 4.8), the data were divided into several files, one for each year; in turn, each year had two files, one containing complete data only, and the other containing all data including data with missing values. It became evident that the accuracy of the adopted models varied from year to year and from one model to another, with the highest accuracy reaching 85% for the year 2015 using the CatBoost model and applying "mean" values to provide the MD. However, no model was able to achieve better accuracy in either scenario. Hence, the role of the third scenario is to improve the accuracy of the models adopted to achieve the highest predictive accuracy, as will be discussed later.

Table 4.8. Model performance using annual data

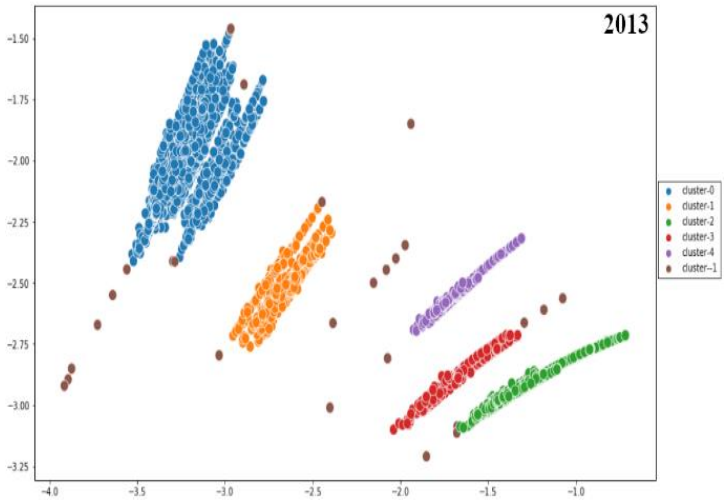
Model	Model Performance				
	Complete records *	Mean	Median	Mode	SOM and DBSCAN
Decision Tree					
2012	65 %	62 %	67 %	65 %	65 %
2013		71 %	71 %	71 %	73 %
2014	55 %	60 %	59 %	61 %	58 %
2015	67 %	76 %	75 %	74 %	69 %
2016	67 %	68 %	67 %	69 %	59 %
2017	65 %	65 %	63 %	67 %	63 %
2018	65 %	67 %	68 %	66 %	62 %
KNN					
2012	58 %	63 %	60 %	58 %	63 %
2013		70 %	69 %	69 %	76 %
2014	60 %	63 %	61 %	64 %	63 %
2015	69 %	71 %	71 %	71 %	71 %
2016	61 %	59 %	60 %	59 %	61 %
2017	59 %	58 %	59 %	59 %	62 %
2018	59 %	61 %	59 %	60 %	65 %
Random Forest					
2012	68 %	72 %	72 %	73 %	66 %
2013		78 %	80 %	78 %	74 %
2014	64 %	68 %	67 %	69 %	58 %
2015	81 %	82 %	80 %	80 %	73 %
2016	74 %	74 %	74 %	73 %	63 %
2017	70 %	71 %	72 %	71 %	62 %
2018	74 %	75 %	75 %	74 %	67 %
AdaBoost					
2012	73 %	75 %	74 %	75 %	70 %
2013		79 %	79 %	78 %	80 %
2014	65 %	70 %	68 %	69 %	65 %
2015	80 %	82 %	80 %	80 %	74 %
2016	68 %	67 %	68 %	68 %	65 %
2017	72 %	70 %	70 %	70 %	64 %
2018	73 %	72 %	73 %	71 %	69 %
SVC					
2012	63 %	63 %	63 %	59 %	54 %
2013		68 %	70 %	69 %	69 %
2014	63 %	65 %	63 %	66 %	65 %
2015	65 %	68 %	67 %	67 %	67 %
2016	63 %	63 %	62 %	62 %	62 %
2017	61 %	62 %	62 %	61 %	54 %
2018	62 %	61 %	60 %	61 %	59 %
CatBoost					
2012	71 %	75 %	74 %	73 %	71 %
2013		80 %	81 %	80 %	81 %
2014	64 %	70 %	69 %	71 %	68 %
2015	83 %	85 %	83 %	83 %	77 %
2016	75 %	77 %	77 %	77 %	67 %
2017	74 %	75 %	76 %	75 %	69 %
2018	76 %	76 %	76 %	74 %	72 %
Naive Bayes					
2012	57 %	67 %	64 %	68 %	56 %
2013		69 %	71 %	69 %	68 %
2014	37 %	35 %	37 %	34 %	60 %
2015	65 %	68 %	66 %	67 %	63 %
2016	37 %	37 %	39 %	38 %	61 %
2017	44 %	45 %	42 %	45 %	47 %
2018	41 %	41 %	41 %	42 %	58 %

* Year 2013 deleted since hospital records are missing for the entire year

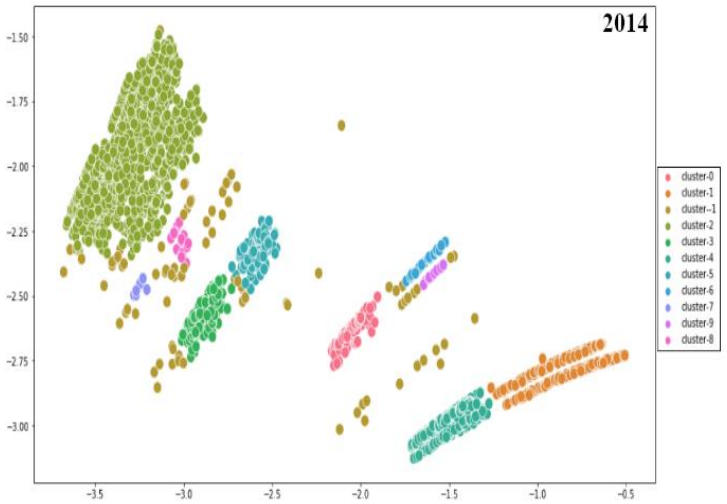
For the last scenario, the imputation of missing values using the clusters generated by SOFM and DBSCAN was different for the numerical and categorical features. To fill in MD for the numerical values related to age, population, temperature, and humidity, mean, median, and mode values were used. Accordingly, a variance table was created to determine the best way to supply the missing value for a particular feature. Apart from the demographic features, all other features can be matched against the collected data to ensure the accuracy of missing values in the variance check which compared the variance of imputed values with the original variance. The lowest variance provides better imputation; hence the imputation of the mean (numerical value) as this produces lower values. Binary encoding and one-hot encoding were applied to features with categorical values to convert these to numerical values. This allows these parameters to be treated as numerical attributes. Before encoding, several features were minimized to produce lower values. For instance, Saudi and Non-Saudi were used for the nationality parameter instead of specifying a nationality. Then, each missing value in the cluster obtained by SOFM and DBSCAN was filled with the most commonly occurring value in that cluster (Figure 4.13). Those patient records that contained a high number of missing values for most of the features were removed from the dataset. After imputing the MD, a prediction model was applied to each cluster individually; Table 4.9 presents the best annual model performance. The clusters' performance annually is given in the appendix (Table A.1). This produced several models that achieved the optimal accuracy of 100% for predicting the disease (cluster 8 for the year 2014) by using a decision tree, AdaBoost, and CatBoost models, while CatBoost achieved the same accuracy for the same cluster using hyperparameters. In 2017, using default values, the KNN model achieved 100% accuracy for cluster 7, while Random Forest and AdaBoost models achieved the same accuracy for the same cluster using hyperparameters.



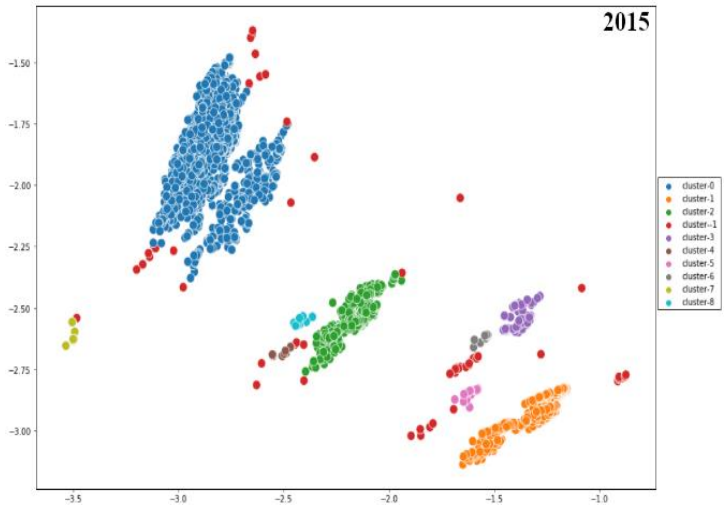
(a) Year 2012 clusters generated by SOFM-DBSCAN



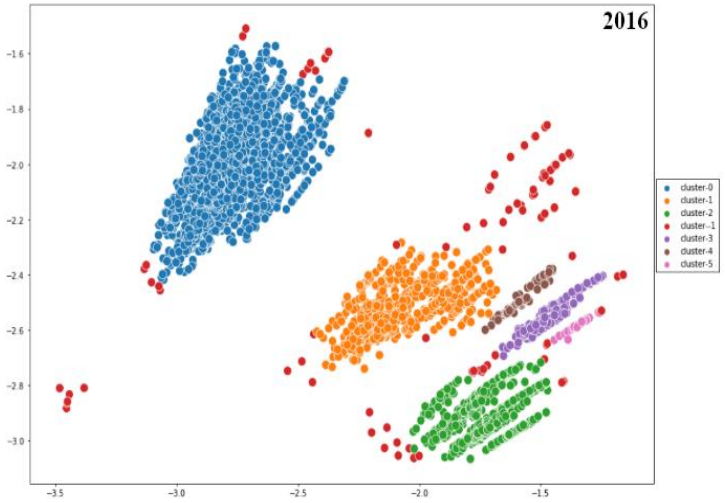
(b) Year 2013 clusters generated by SOFM-DBSCAN



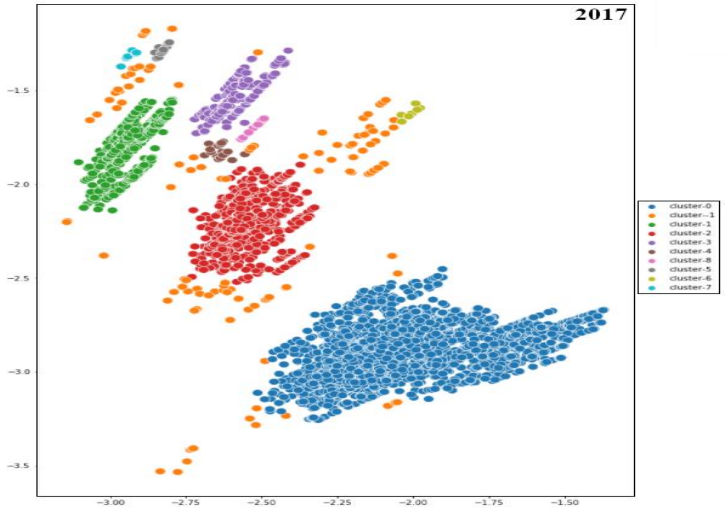
(c) Year 2014 clusters generated by SOFM-DBSCAN



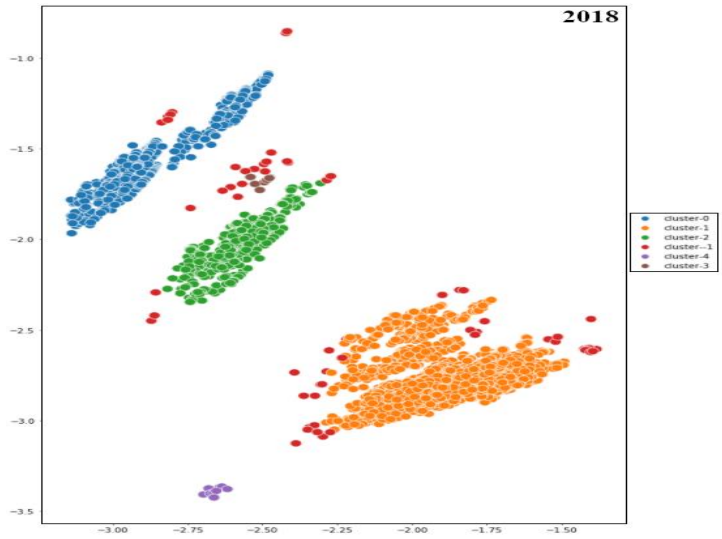
(d) Year 2015 clusters generated by SOFM-DBSCAN



(e) Year 2016 clusters generated by SOFM-DBSCAN



(f) Year 2017 clusters generated by SOFM-DBSCAN



(g) Year 2018 clusters generated by SOFM-DBSCAN

Figure 4.13. Yearly clusters generated by SOFM-DBSCAN

Table 4.9. Model performance using annual clusters data

Year	Cluster	Model	Accuracy
2012	4	KNN, Naive Bayes	83 %
2013	1	AdaBoost, CatBoost	83 %
2014	8	Decision Tree, Random Forest, AdaBoost, CatBoost	100 %
2015	5	AdaBoost	83 %
2016	5	KNN, Random Forest, AdaBoost	82 %
2017	7	Decision Tree, KNN, Random Forest, AdaBoost, SVC,	100 %
2018	4	Decision Tree, Random Forest, AdaBoost, SVC, Naive Bayes	86 %

4.4.2 Discussion

MD is a major issue in most research and affects the quality of the collected data (Batista and Monard 2003). Hence, over the last few decades, MD has been a popular research topic in data mining, and various methods have been proposed for dealing with this issue (Sessa and Syed 2016). The prevalence and importance of the MD problem has prompted researchers to analyse it from different perspectives and to investigate optimal

methods for handling it (Brown and Kros 2003; García-Laencina et al. 2010; Jakobsen et al. 2017; Kang 2013; Lazar et al. 2017; Ngueilbaye et al. 2021; Pigott 2001; Scheffer 2002; Sessa and Syed 2016; Zhang 2016). To ensure accuracy, the development of a robust data analytical model is required for handling MD imputation (Ngueilbaye et al. 2021). Furthermore, MD imputation should be dealt with appropriately to avoid bias in the study results (Batista and Monard 2003). Generally, the accuracy of any approach used to estimate MD is strongly influenced by the patterns and relationships evident in the data, and the way in which the data is missing (Richman et al. 2009). Therefore, the current study applied several imputation strategies to three scenarios and evaluated their impact on various models.

Several previous studies used the same models for the prediction modelling of DF and achieved various levels of accuracy based on the dataset and the applied framework. In alignment with the accuracy of our findings (Sajana et al. 2018), Sajana conducted a comparison study that achieved the optimal accuracy of 100% when classifying confirmed cases and distinguishing them from unaffected patients using simple classification and regression tree (CART) as well as other Multi-Layer Perception (MLP) and C4.5 algorithms. Another study (Faisal et al. 2010) combined SOM and multilayer feed-forward neural networks (MFNN) to classify the dengue cases, achieving a prediction accuracy of 70%. Another paper modelled the severity of DF in 1,225 children's records based on clinical and laboratory features, using the decision tree algorithm to achieve 64.1% accuracy (Phakhounthong et al. 2018). Lastly, decision tree (DT) and random forest (RF) algorithms were applied to data on 209 patients to classify three types of DF and achieved 79% accuracy using the DT (Sarma et al. 2020). However, in previous studies, most researchers ignored the impact of data quality on the validity of results; nor did any propose a comprehensive framework that delivers better performance.

The main contribution of the current study is the improvement of DF transmission prediction performance in the presence of MD. The proposed model is transferable to other datasets with missing values issues. Following previous works, the same approach to improve the model's performance used for the prediction of breast cancer survivability was adopted (Shukla et al. 2018). However, by adopting the same approaches for the imputation of missing values, and modelling each cluster individually, the current study applied several imputation strategies and modelling algorithms and compared the impact of each method on the final performance. By doing so, the current study achieved the better accuracy of 100% for predicting dengue cases for some clusters.

In this study, four imputation strategies were applied to handle the problem of missing values in Saudi's dengue data. Initially, the adopted models were applied to a complete data set obtained by discarding all records with missing values for at least one feature. However, this method was not applicable here as missing values were found in over 5% of the collected records (Jakobsen et al. 2017), and was applied to analyse the model performance on data that had no missing values. Following the previous techniques, three statistical imputation methods (mode, mean, and median) were applied to fill in the missing values and were then compared to advanced machine-learning-based approaches by grouping the patients' records into several clusters generated by SOFM-DBSCAN. These MD imputation algorithms were then applied to the dataset of dengue disease patients, 24.7% of which had missing values. Table 4.6 shows the features with missing values in records from 2012 to 2018. Following the imputing of missing values, several imputation methods were applied to test the performance of models used to predict DF cases and to compare the effectiveness of different imputation methods. Dengue disease prediction is an important means of determining the risk areas and associated risk factors; thus, a key aim of this study was to determine the imputation techniques that

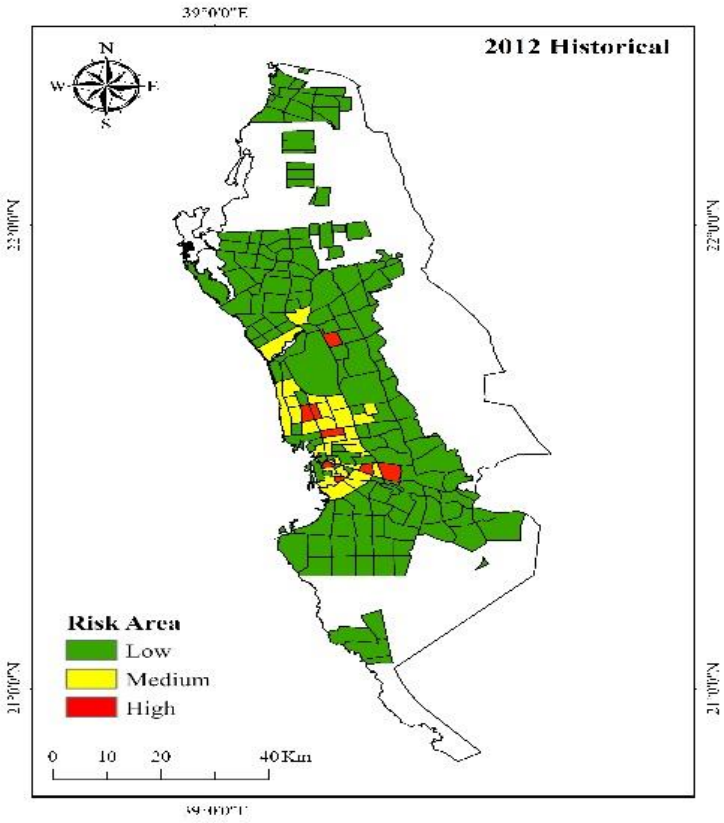
could improve the prediction accuracy. Adopted imputation techniques were applied to three scenarios (all data, yearly data, and cluster data), yielding different levels of prediction accuracy, measured by the confusion matrix. For the purpose of comparison, Tables 4.7, 4.8 and 4.9 show the results obtained by SOFM-DBSCAN and those of other approaches, and indicate any discrepancies. The differences in results are significant when the proposed approach is compared with the others, as it achieves 100% prediction accuracy for some clusters.

Despite limitations of this objective, a novel approach has been developed for data imputation that is intended to help the research community concerned with data quality, particularly with regard to the imputation of MD. Moreover, the developed approach has demonstrated several advantages. First, by observing the relative prediction accuracy, it is anticipated that the proposed imputation technique will produce superior accuracy than that achieved by the traditional approaches. Second, this is the first study to apply dengue prediction modelling to investigate the impact of missing values in the final models and to propose an optimal framework to improve the prediction accuracy. Third, this study addresses a gap in the research by demonstrating that the separate modelling for each cluster generated by SOFM-DBSCAN can improve prediction accuracy and give a better understanding of the factors associated with the disease in each cluster. Finally, despite the complexity of the proposed imputation techniques, the methodological framework is valuable and transferable to other datasets.

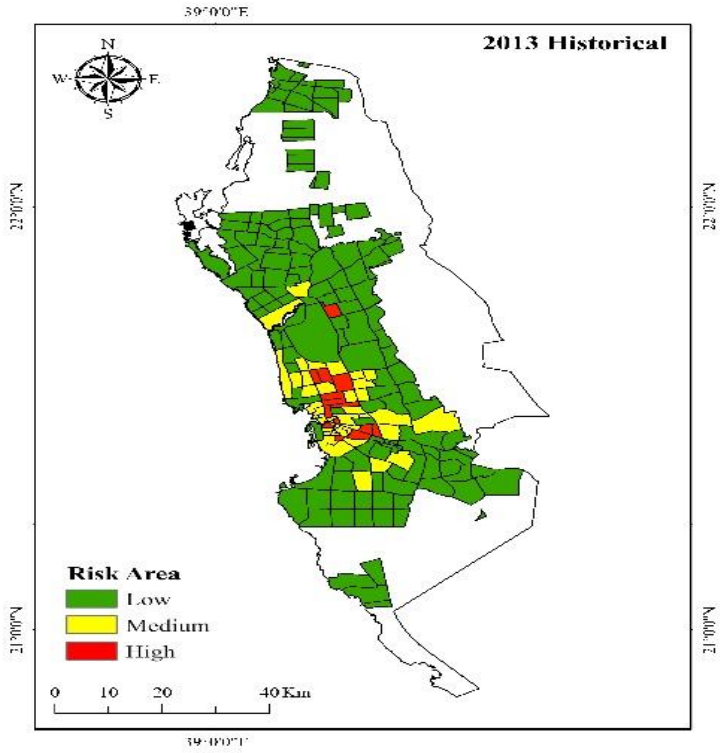
4.5 Objective 3 (Simulate risk areas)

Analysis revealed that the number of reported cases is dynamic and changes annually, although those at highest risk are often concentrated in the central neighbourhoods of a region. The MOLUSCE-plugin is used only for rasters to prepare

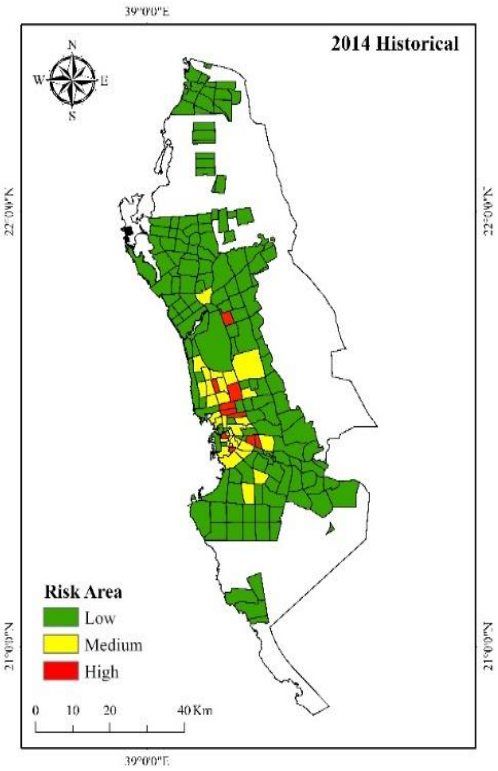
them for simulation, not for other formats. Thus, all input variables were unified in terms of the number of classes, coordinated system, and pixel size to match the geometries of all rasters. This initial step is essential prior to conducting subsequent steps. Firstly, the efficiency of MOLUSCE-plugin was validated using the total average of each variable to predict the 2018 risk map; annual statistics for each year were used to simulate the risk of the following year. Then the simulated map was compared with the actual map of the risk area. For instance, the 2012 risk map was used as the initial layer and the 2013 map was the final layer; the simulation iteration was set to 1 (one) to simulate the map for the following year. In this case, the risk map for 2014 was generated. This process was repeated to cover the annual data for 2014 to 2018. The actual risk areas were compared with the simulated maps, except for the 2012 and 2013 maps as there were no data for 2010 and 2011, as Figure 4.14 illustrates. Results indicate that the CA is capable of correctly simulating most of the districts at risk.



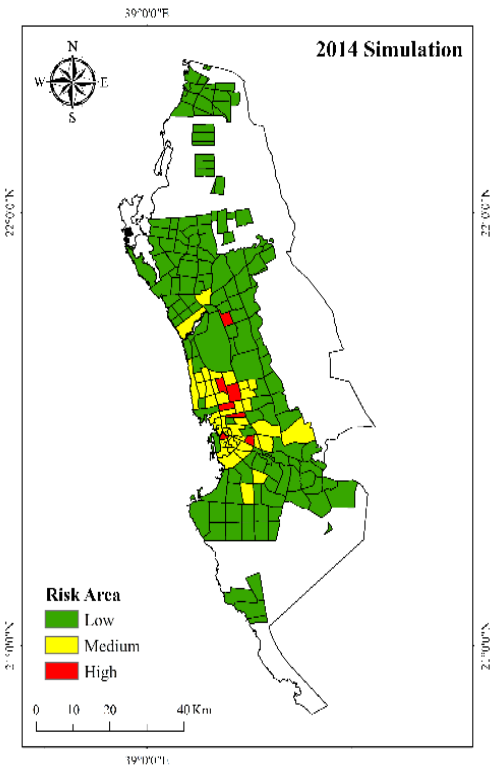
(a) 2012 historical risk areas



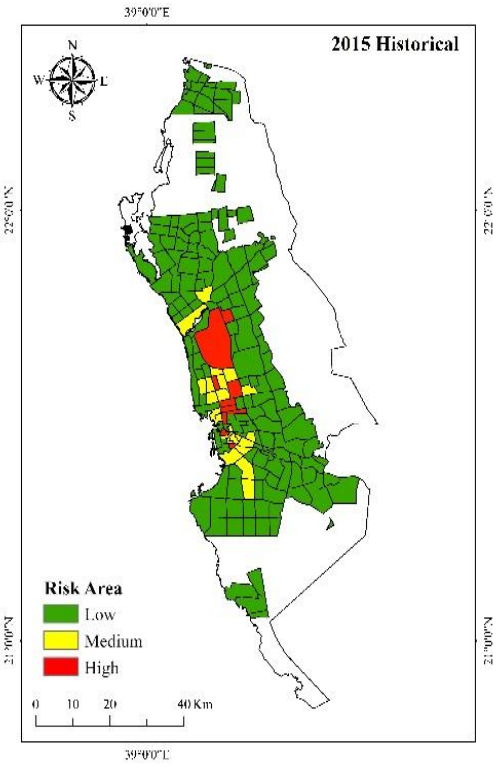
(b) 2013 historical risk areas



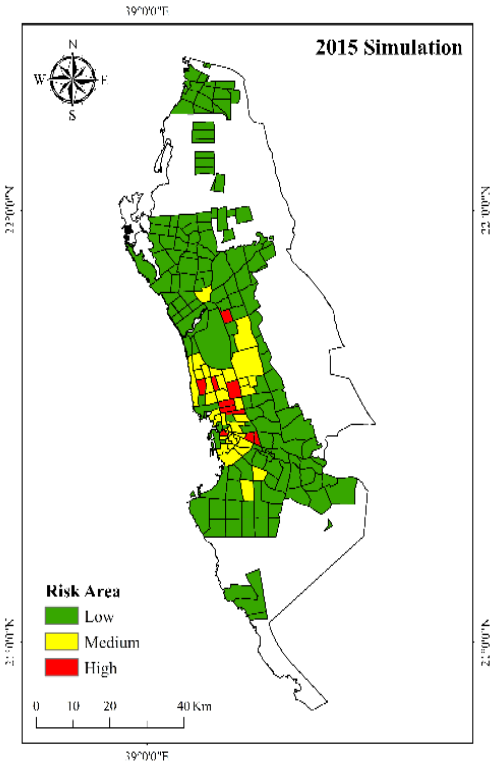
(c) 2014 historical risk areas



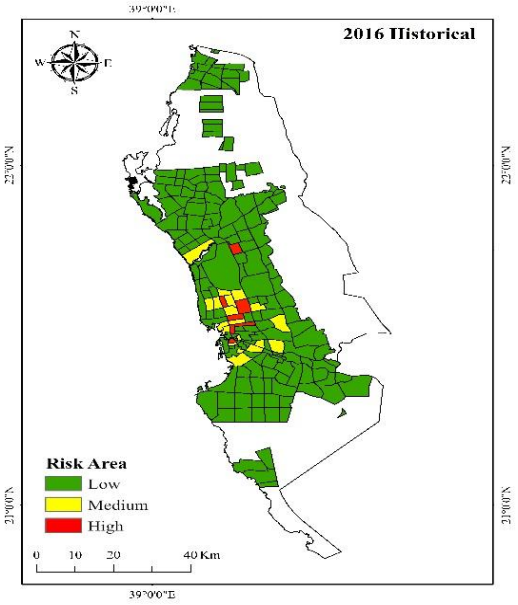
(d) 2014 simulated risk areas



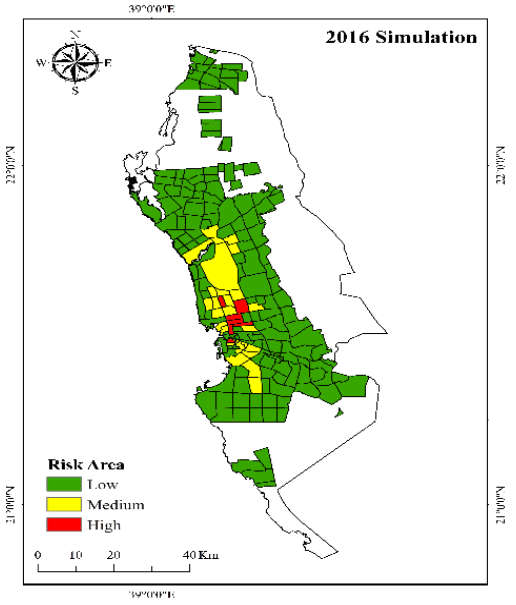
(e) 2015 historical risk areas



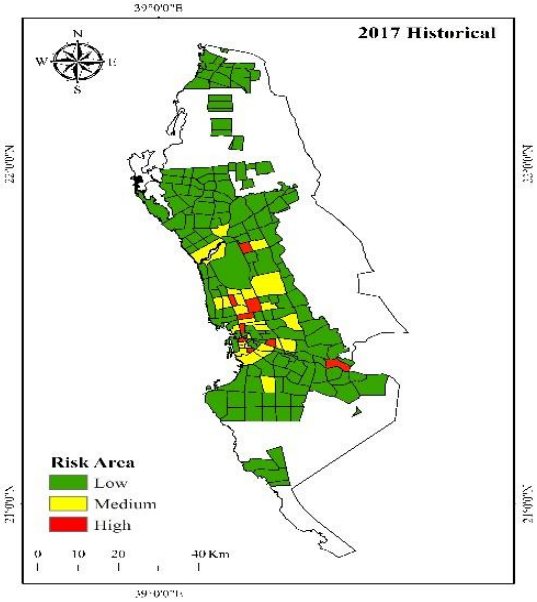
(f) 2015 simulated risk areas



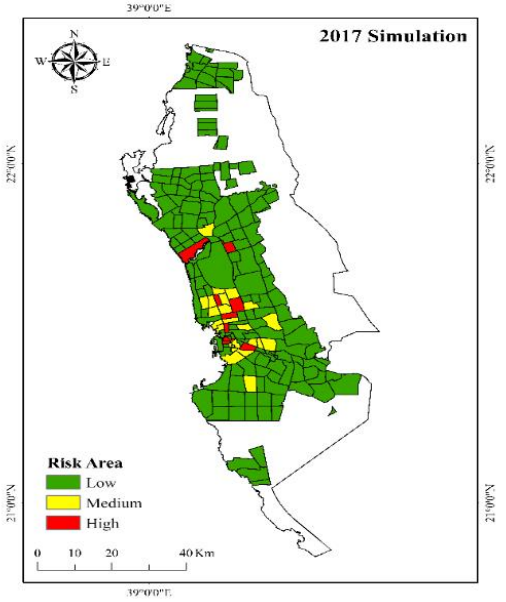
(g) 2016 historical risk areas



(h) 2016 simulated risk areas



(i) 2017 historical risk areas



(j) 2017 simulated risk areas

Figure 4.14. Annual historical "Actual" vs. simulated risk maps; (a) 2012 historical risk areas, (b) 2013 historical risk areas, (c) 2014 historical risk areas, (d) 2014 simulated risk areas, (e) 2015 historical risk areas, (f) 2015 simulated risk areas, (g) 2016 historical risk areas, (h) 2016 simulated risk areas, (i) 2017 historical risk areas, and (j) 2017 simulated risk areas

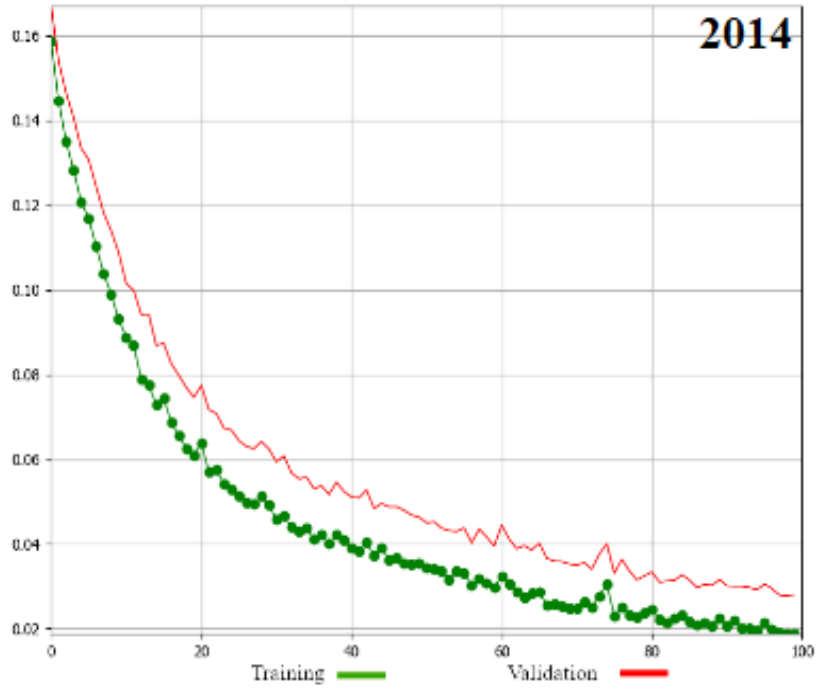
4.5.1 Results of cellular automata model

After evaluating the efficiency of MOLUSCE tool as explained in the previous paragraph, each scenario was examined. After matching the geometries of all rasters, in the second step, the correlation of these factors with the number of cases was investigated. The MOLUSCE tool provides three algorithms for this purpose: Pearson's correlation coefficient, Cramer's correlation coefficient, and joint information uncertainty. Pearson's correlation coefficient was computed to investigate the relationship between spatial variables as these factors are not categorical data. Pearson's correlation coefficient measures the extent to which variables are related. Values close to 0 indicate a weak relationship between two variables, while values closer to 1 and -1 indicate a strong relationship between the variables. This approach was applied to the annual statistics to determine the correlation between the number of cases and the significant factors being investigated. The appendix gives details of the correlation findings for each year (Table B (1-5). Table 4.10 below gives only the range of correlation values between the annual confirmed cases and the related spatial variables. Based on the correlation values, population density is the most significant factor influencing the annual number of confirmed cases, as the Pearson's correlation is between 0.46 and 0.52.

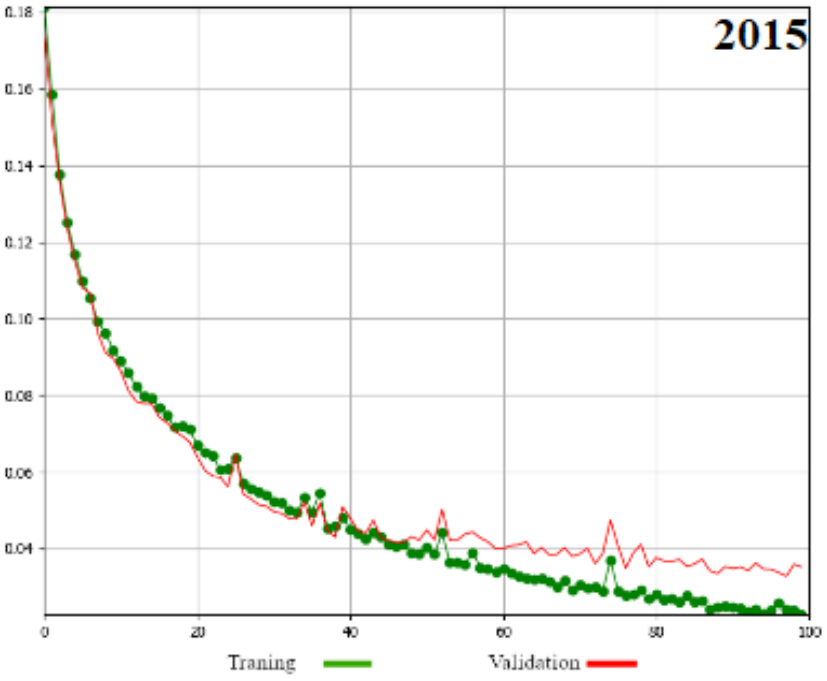
Table 4.10. Pearson's correlation values for dependent and independent parameters (annual statistics)

	Cases	Humidity	Elevation	Wind Speed	Precipitation	Temperature	Population	LULC
Cases	-	0.20 ≥ 0.34	-0.24 ≥ -0.05	0.06 ≥ 0.18	-0.20 ≥-0.04	-0.05 ≥ 0.07	0.52 ≤ 0.46	-0.55 ≥ -0.41
Humidity		-	-0.54 ≥ -0.41	-0.14 ≥ 0.09	-0.27 ≥-0.01	-0.07 ≥ 0.09	0.17 ≥ 0.24	-0.41 ≥- 0.38
Elevation			-	0.46 ≤ 0.26	0.15 ≤ 0.03	0.03 ≥ 0.31	-0.16 ≥ -0.12	0.19
Wind Speed				-	-0.07 ≥ 0.17	0.19 ≥ 0.71	-0.01 ≥ 0.09	-0.24 ≥ -0.12
Precipitation					-	-0.02 ≥ 0.11	-0.07 ≥ 0.002	-0.11 ≥ 0.1
Temperature						-	-0.04 ≥ 0.02	-0.14 ≥ 0.07
Population							-	-0.54 ≥ -0.44
LULC								-

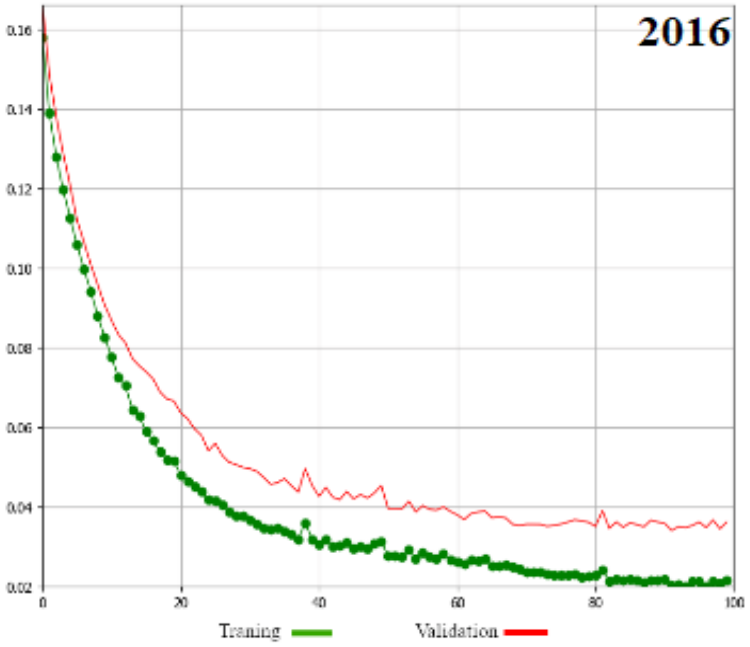
In the third step, the potential transition map was calculated using the MOLUSCE plugin which provides numerous algorithms for computing the transitional potential map, including Artificial Neural Network (Multi-layer Perceptron), weights of evidence, multi-criteria evaluation, and logistic regression. In this study, MLP-ANN was used as it has robust properties, as mentioned in the previous section. This method is appropriate when the algorithm is required to cope with massive volumes of unknown or difficult-to-implement input data (Kamaraj and Rangarajan 2022). Therefore, the parameters were adjusted to fit the model using the collected data to produce the best values. To assess the MLP, the random sampling technique was applied using 2000 points as the sample size. The other parameters were set as follows: neighbourhood is 1 pixel, learning rate and Momentum are 0.001, iteration is 100, and hidden layer is 7. Figure 4.15 depicts the MLP-Neural Network learning curve for modelling changes in annual case numbers. Table B (1-5) in the appendix shows the annual Neural Network learning inputs in detail.



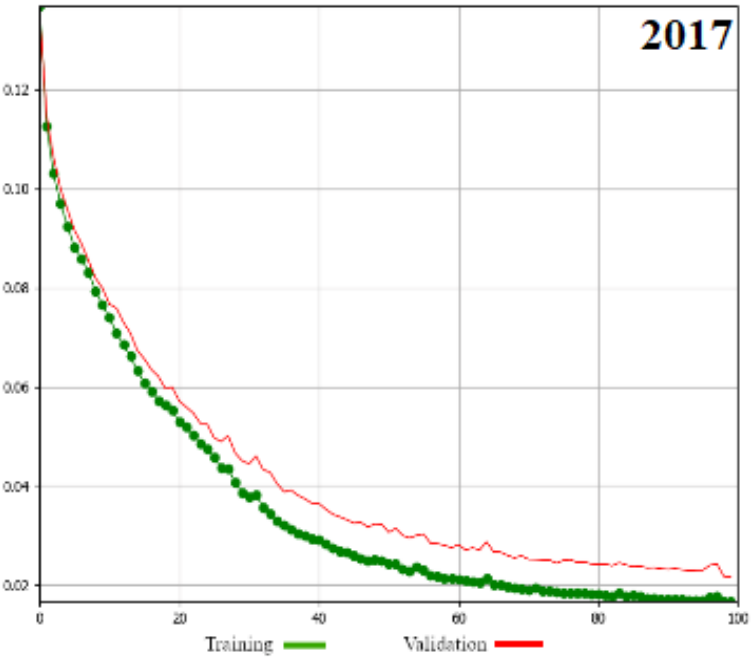
(a) year 2014 learning curve



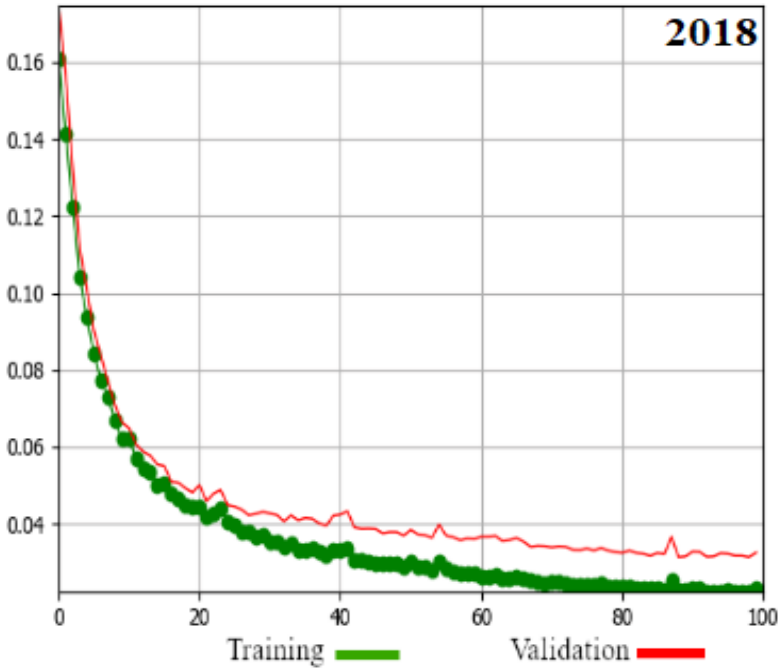
(b) year 2015 learning curve



(c) year 2016 learning curve



(d) year 2017 learning curve



(e) year 2018 learning curve

Figure 4.15. Neural network learning curve; (a) year 2014 learning curve, (b) year 2015 learning curve, (c) year 2016 learning curve, (d) year 2017 learning curve, and (e) year 2018 learning curve

The penultimate step prior to assessing the prediction map, involves determining the year to be predicted by specifying the number of simulation iterations. For the annual assessment map for the first scenario, the number of iterations was set to 1 (one) to generate one year ahead since the temporal interval between the first and final layer is one year. The overall prediction accuracy ranged between 88.78% and 92.52%, as the 2017 map has the highest accuracy and the 2015 map is the least accurate of the prediction maps with Kappa values ranging between 0.48 and 0.65, as shown in Table 4.11 below.

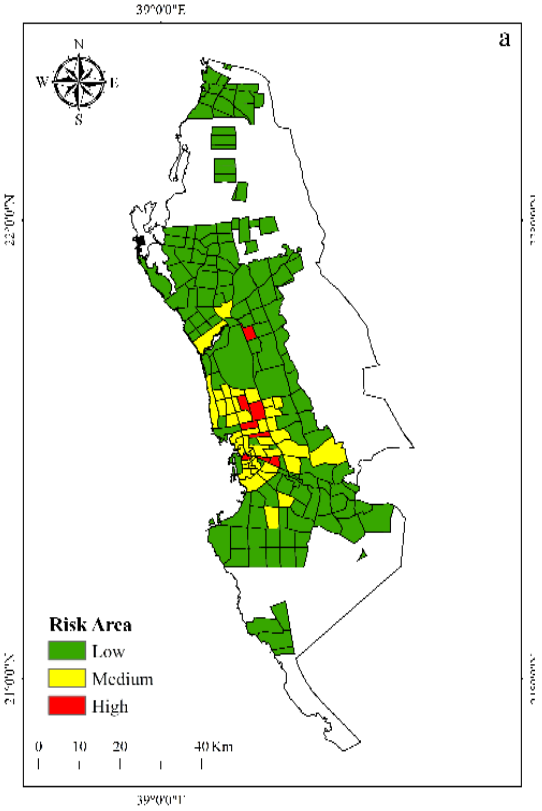
Table 4.11. Simulated map evaluation parameters

Parameters	2012	2013	2014	2015	2016	2017
	Values (%)					
% of correctness	-	-	90.14 %	88.78 %	91.57 %	92.52 %
Kappa (overall)	-	-	0.61	0.48	0.59	0.65
Kappa (histogram)	-	-	0.86	0.94	0.78	0.78
Kappa (location)	-	-	0.71	0.51	0.76	0.83
MLP Validation Kappa	-	-	0.86	0.94	0.87	0.95

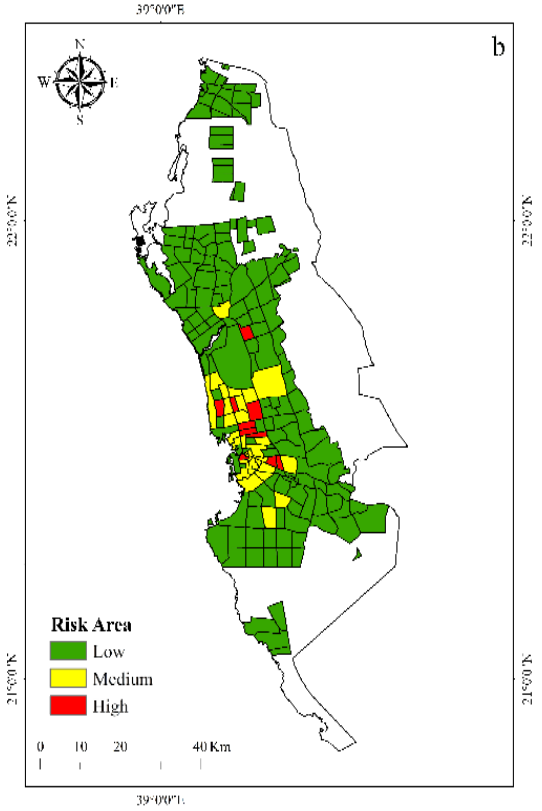
The past trends of confirmed disease data were used in the ANN model to predict future disease trends for 2018. Thus, in the second scenario, the risk map for 2018 was simulated based on the annual number of cases. The number of simulated iterations was changed accordingly; then, the actual 2018 map was used as a reference to determine the accuracy of the simulated map. Table 4.12 shows the outcomes of the Kappa values in addition to Kappa validation values for the simulated 2018 maps using annual cases maps. Although the Kappa values of 0.47 and 0.71 indicated substantial and moderate agreement respectively, the validation yielded an outstanding accuracy of over 87% for the 2018 simulated risk map. Moreover, using 2016 and 2017 maps as inputs to simulate the 2018 map, the latter performed better than in the previous years in terms of map accuracy and Kappa values. Figure 4.16 below shows the simulated map compared with the actual 2018 map. Figure 4.16 (a) shows the simulated risk map using map 2012 as the input map. Figure 4.16 (e) is based on the 2016 risk map. The model validation results shown in Figure 4.17 were obtained using the QGIS program.

Table 4.12. Assessment of simulated 2018 map based annually

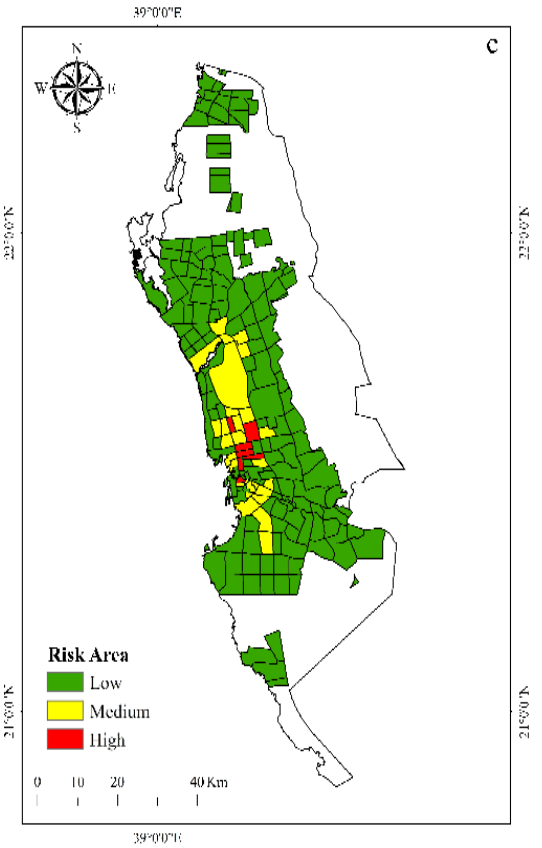
#	Initial	Final	Iterations	% of correctness	Kappa (overall)	Kappa (histogram)	Kappa (location)	MLP Validation Kappa
1	2012	2013	5	89.81	0.62	0.90	0.68	0.92
2	2013	2014	4	90.69	0.60	0.90	0.67	0.92
3	2014	2015	3	87.59	0.47	0.94	0.50	0.89
4	2015	2016	2	91.33	0.60	0.78	0.76	0.93
5	2016	2017	1	92.76	0.71	0.97	0.83	0.92



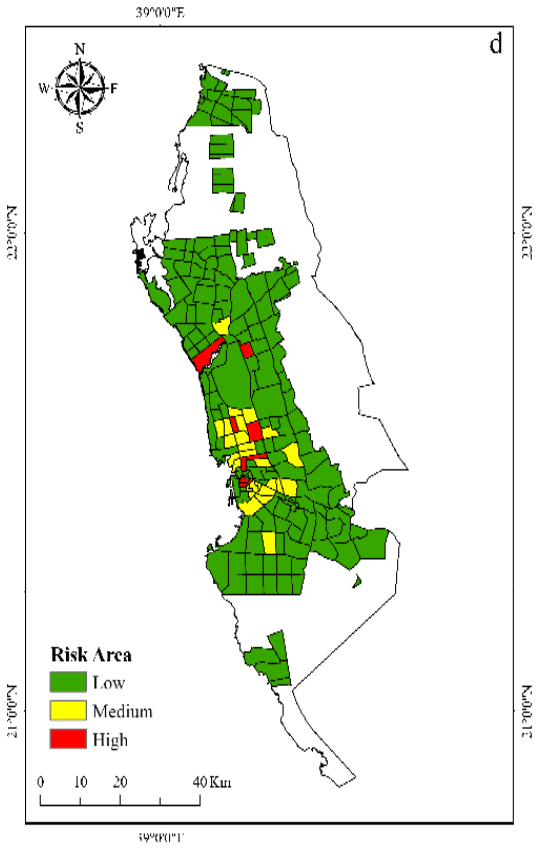
(a) simulate year 2018 risk map using 2012 historical data



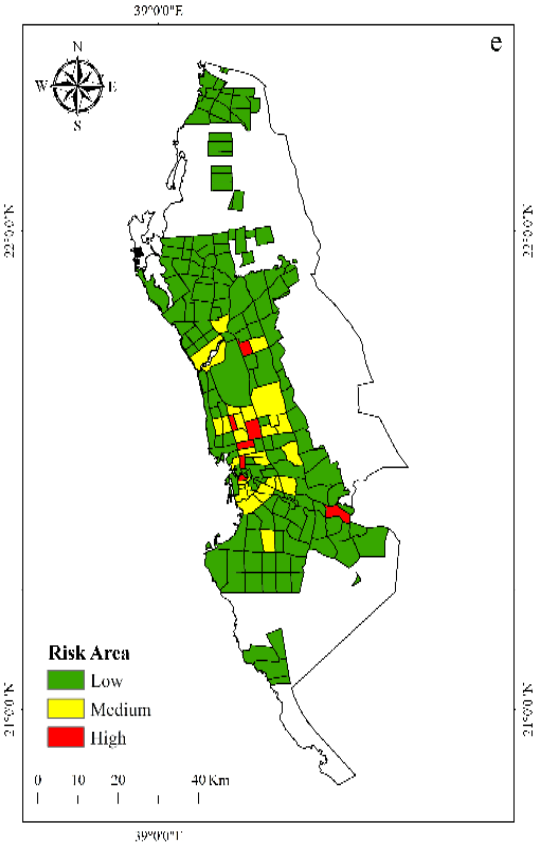
(b) simulate year 2018 risk map using 2013 historical data



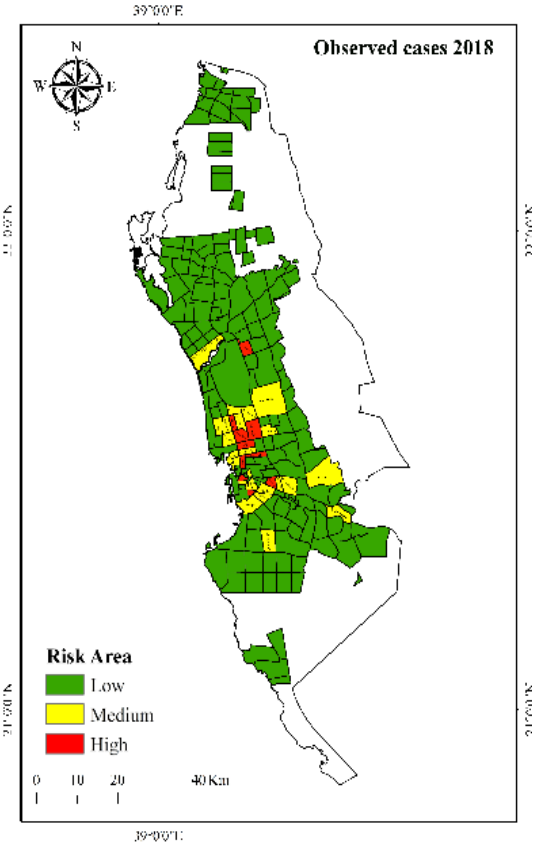
(c) simulate year 2018 risk map using 2014 historical data



(d) simulate year 2018 risk map using 2015 historical data



(e) simulate year 2018 risk map using 2016 historical data



(f) year 2018 historical risk map

Figure 4.16. Actual 2018 risk map vs. simulated map; (a) simulate year 2018 risk map using 2012 historical data, (b) simulate year 2018 risk map using 2013 historical data, (c) simulate year 2018 risk map using 2014 historical data, (d) simulate year 2018 risk map using 2015 historical data, (e) simulate year 2018 risk map using 2016 historical data, and (f) year 2018 historical risk map

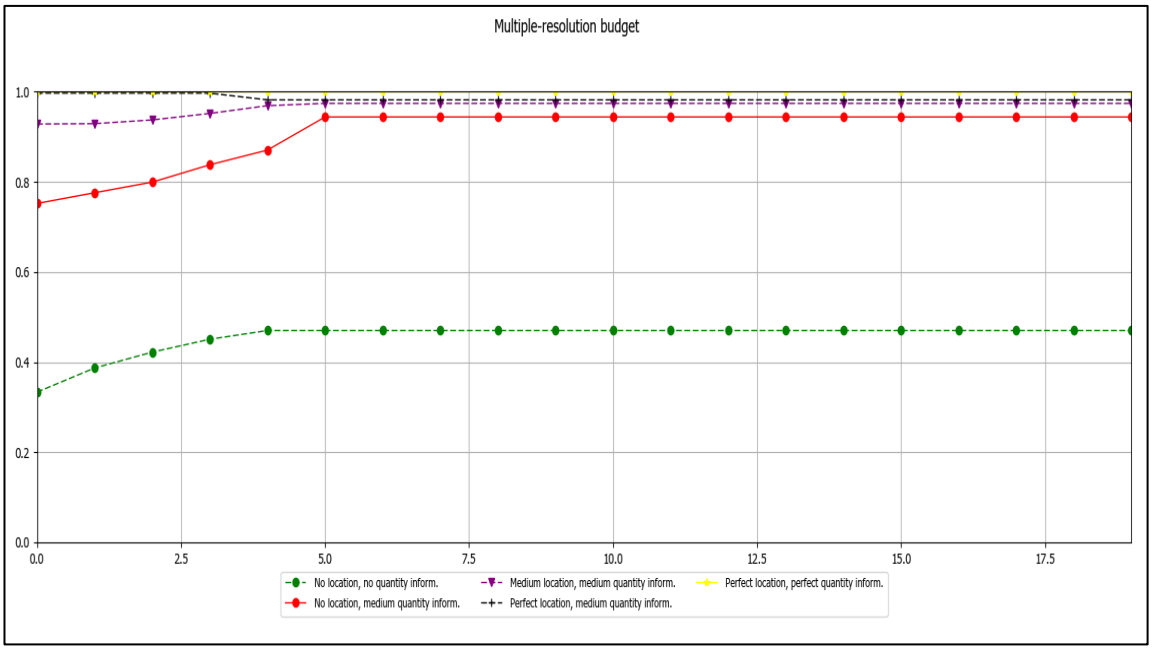
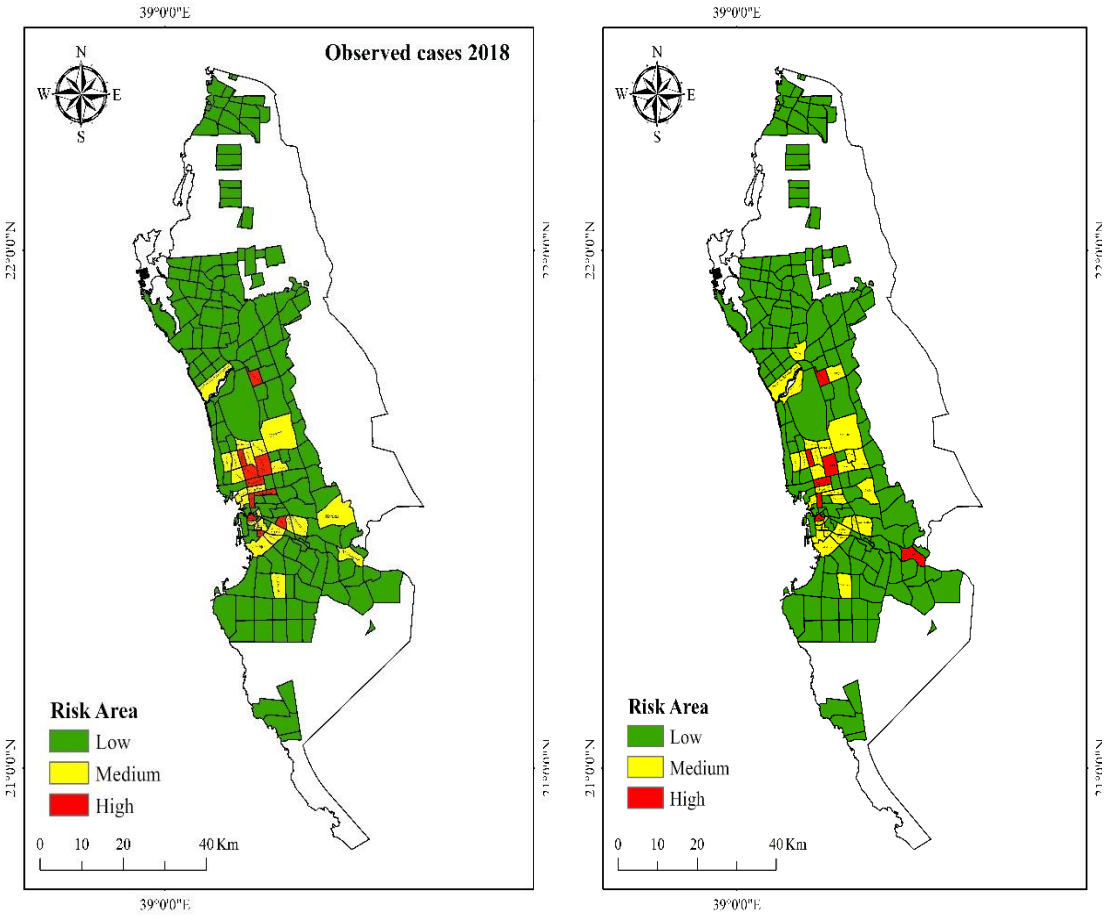


Figure 4.17. Validation graph showing actual 2018 map and 2018 simulation map predicting dengue cases

Upon comparing the risk areas in the observed data with the simulated map for 2018, it is clear that most of the notified cases were in the central districts. The observed data show that high-risk areas are Al Hamadaniyyah, Al Bawadi, Al Faisaliyyah, Bani Malik, Al Sharafeyah, Al-Balad, Ghulail, Al Jami`ah, Al-Safa, and Aziziyah. The simulated map shows most of the high-risk districts that are also in the observed data in addition to Al Mursalat area, while four neighbourhoods are simulated as medium-risk districts: Al Faisaliyyah, Bani Malik, Ghulail, Al Jami`ah (Figure 4.18). In addition to the four previous districts simulated as medium-risk areas, ten other districts are in this category. However, according to both maps, eighteen districts had similar levels of risk except for Ar Rawdah, Al-Hamra'a, Al Kandarah, Al-Nazlah Al-Yamaniyah, Harazat, and Al Mursalat which were in the recorded data.



(a) (b)

Figure 4.18. Observed vs. simulated 2018 risk map; (a) year 2018 risk map based on historical data, and (b) year 2018 simulated risk map

In the third scenario, in addition to investigating the impact of the data on the simulation process, the average of each factor for the entire study period was calculated using "Cell Statistics" in ArcMap. Following the same simulation stages, it is obvious that simulated high/medium/low risk areas are the exact same when they are simulated using the average of each parameter as the spatial variable and the year 2012 cases as initial input and year 2017 as the final layer after adjusting the simulation to 1 (one) iteration as shown in Figure 4.19. Setting the simulation iteration to 1 (one) is intended to simulate the 2022 map because the gap between the initial and final layer is five years.

However, the average values were considered as one year and the same map was simulated using 2016 and 2017 as input rasters. Table 4.13 shows the CA validation parameters of the 2018 simulation based on average values, using 2016 and 2017 inputs for the simulation of the 2018 reference map. Assuming that the predictions for 2018 are aligned with the recorded data, it can be concluded that predictions for other years in the future will also be valid.

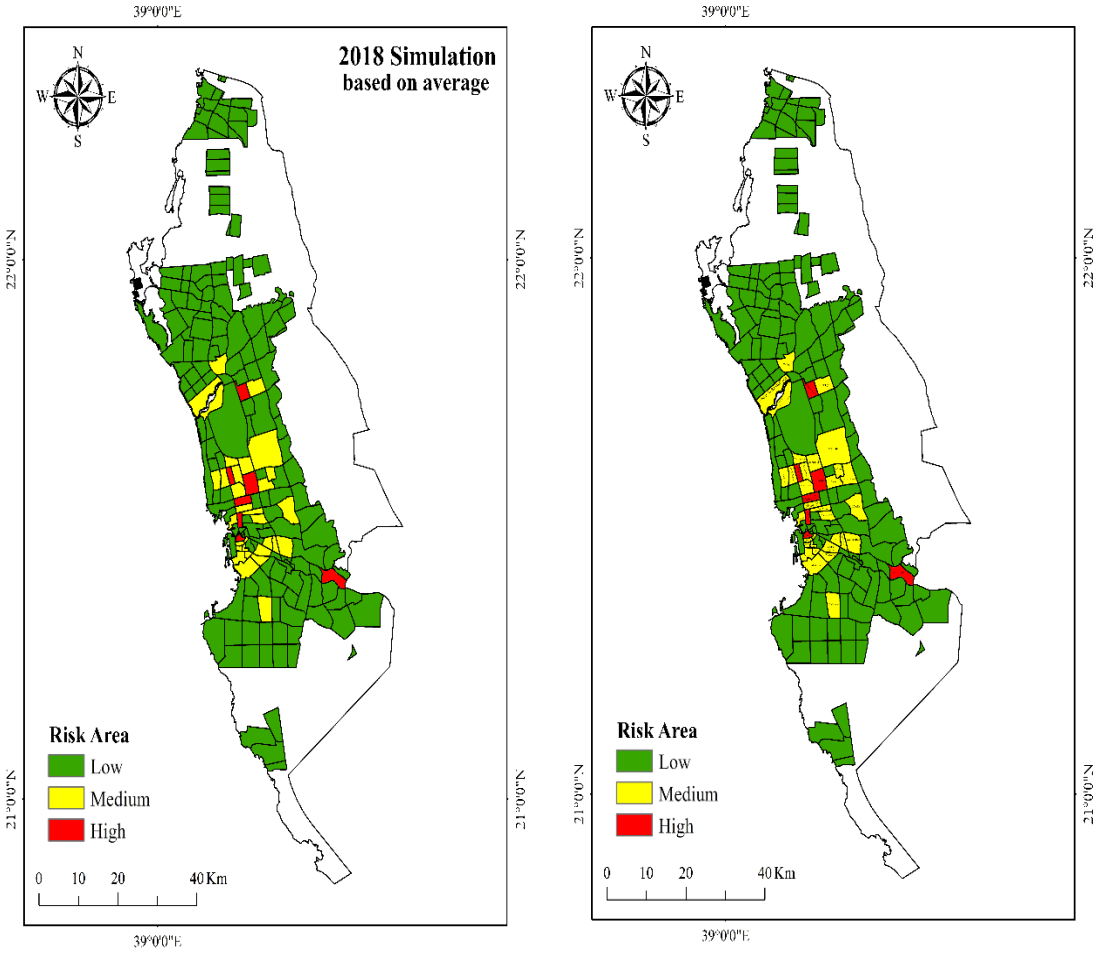


Figure 4.19. Simulated year 2018 risk maps using two datasets; (a) simulated 2018 cases based on average values for all factors (2012-2016), (b) simulated 2018 cases based on year 2016-2017 data

Table 4.13. Parameter values of the model validation (simulated 2018 cases based on average values for all years 2012-2017 vs. simulated 2018 cases based on year 2016-2017 as reference)

Parameters	Values (%)
% of correctness	96.34 %
Kappa (overall)	0.84
Kappa (histogram)	0.87
Kappa (location)	0.97

4.5.2 Discussion

Computational modelling improves decision-maker's understanding of the temporal and spatial dynamics of epidemics, as well as guiding, testing, and modifying control tactics in simulations prior to their implementation in the real-world (Lemos et al. 2017). Various computer modelling approaches have been utilized to model, simulate, and subsequently comprehend different social, environmental, biological, and other types of systems (Khalil and Wainer 2020).

Numerous studies have utilized the CA model because it is dynamic, can be integrated with other models, and can be modified to suit the available data (Falah et al. 2020). A work by (Ortigoza et al. 2019) focuses mainly on CA as it has been applied in previous works to model the transmission of vector-borne diseases caused by mosquitos. In their paper, the pros and cons of previous models are examined and discussed; also, the models are classified according to the available data. Their findings show that CA is the best approach for conducting simulations of mosquito-borne diseases as it allows for flexibility when examining spatial and temporal patterns. Furthermore, (Eosina et al. 2016) used CA in their study to provide a prediction model and visualize the transmission patterns of dengue haemorrhagic fever, introducing a new method for creating a probabilistic function that captures the CA transmission rule using the Von Neumann neighbourhood and the Hidden Markov Model (HMM). The authors found that the CA

model is capable of creating patterns that are comparable to those formed by Susceptible-Infected-Recovered (SIR) models, with a similarity value of 0.95 approaches. In other research, (Pereira and Schimit 2018) studied the interaction between humans and dengue-causing mosquitoes were simulated using CA. Their proposed framework is capable of capturing the different dynamicity of the disease transmission for two or three dengue serotypes simultaneously; moreover, the model is flexible enough to be applied to other vector-borne diseases.

In another work, to improve the accuracy of predictions, the Artificial Neural Network (ANN) approach was applied using historical dengue cases data along with average temperature and rainfall as highly-correlated factors for disease transmission in the Visayas region of the Philippines (Datoc et al. 2016). Focusing on a single serotype (DENV-1), Enduri and Jolad (2018), predicted the disease transmission by investigating the human mobility and the mosquito density as a proxy for weather data as input in the simulation process. Their results show that the epidemic appears to be suppressed sooner than in stationary individuals, even though the human movement speeds up the transmission of the disease.

The extinction of disease was investigated by researchers who applied the CA algorithm and used birth, death rate and migration as input variables (Sun et al. 2010). They found that the disease disappears for a while in a single patch when the infection rate is low or large enough. However, when the invasion is simultaneously both ‘stable spiral’ and ‘turbulent wave’, the disease will persist. By simulating the impact of different parameters related to humans and the vectors transmitting the virus for extended periods (Medeiros et al. 2011), the researchers found that dengue transmission is prevented by small human populations and low renewal rates. Additionally, the finding showed that the virus can circulate low values of house index for extended periods. In order to

determine the regions in Nicaragua with a high risk of dengue, and to identify the factors contributing to this risk, (Theodorakos et al. 2017) applied a SIR predictive model in cellular automata. SIR is comprised of three layers: host (human), vectors and environment/climate within gridded Meta populations. These researchers established Moore neighbourhoods of varying sizes within which there was host movement. A training method based on differential evolution was applied to fit the time-series data of the geo-referenced host/vector population. Moreover, they found that the "most important fitted variables were: urban/rural land classification, altitude and rainfall time series. The CA makes it possible to identify local effectors based on detailed spatial data and transmission dynamics.

In separate work, (Santos et al. 2009) used the CA comprising three vector population levels to reduce the factors to those that are susceptible to seasonal fluctuation (human, adult and immature vector populations). External seasonal factors that trigger human and mosquito migration, and vector control measures are some of the parameters applied to model the spatiotemporal transmission of the disease. Moreover, data on dengue epidemics in two cities in Brazil are compared in order to derive the results. The results show the variations that mean-field models do not capture. They also reveal the qualitative behaviour of the epidemics' spatiotemporal patterns. In the absence of an external periodic drive, the model predicts a very distinct long-term evolution. The model is robust enough to reproduce the time series of dengue epidemics in different cities, provided that the forcing term considers any changes in the local rainfall.

In Jeddah, many dengue disease cases are unmonitored due to the complex interaction of several climatic, socio-economic, and environmental factors. Moreover, there has been a continuous increase in the number of cases of dengue reported in recent years. Additionally, there is a lack of studies on spatiotemporal patterns and clusters of

dengue cases at the district scale for the study area. Thus, the present study attempts to fill this gap by predicting the DF cases based on historical reported cases and validating the model performance using the year 2018 DF classified image. Moreover, assess the feasibility of using artificial neural networks and cellular automata to simulate and model DF. The changes in disease cases from 2012 to 2017 were detected by an Artificial Neural Network (ANN) analysis, and the prediction of disease threats for 2018 is done through cellular automata. In addition to predicting the future scenario of DF cases in 2018, this study will apply the Cellular Automata Model by adopting the MOLUSCE Plugin in QGIS software. By filling the knowledge gap, we can increase our understanding of the spatiotemporal pattern of DF and allocate resources to prevent and control it more precisely.

In here, the prediction of cases of DF disease based on CA method was carried out using the MOLUSCE plugin. To the best to the author's knowledge, this study is the first attempt to simulate DF cases using the QGIS/MOLUSCE plugin, as most of the previous studies have used it to simulate LULC changes (Al Kafy et al. 2021; Aneesha Satya et al. 2020; Baidya et al. 2021; Ferdous and Rahman 2019; Jogun et al. 2019; Kafy et al. 2021; Kamaraj and Rangarajan 2022; Kositsakulchai et al. 2021; Lakshita and Rahayu 2021; Rahman and Rahman 2021; Rahman et al. 2017; Saha et al. 2021; Ullah et al. 2019). However, the observed data were adjusted to fit the tool as well as the prediction model. Pearson's correlation indicates that population is the main factor accounting for changes in the annual number of cases. Transitional changes in case maps from one year to the next were simulated using an artificial neural network (ANN), while cellular automata simulation was conducted to predict potential future cases for the year 2018 after validating the model using the risk map for cases recorded in 2018. In order for the simulation to produce valid results, the reference data must be accurate and comparisons

between various models are relative (Jogun et al. 2019). Thus, a proportion of accurate predictions was used to assess prediction accuracy (% of correctness), the overall Kappa value, the Kappa history value, and the Kappa location value for the predicted variables. The model validation findings for transition potential modelling indicated that the correctness percent value was above 92% and the overall Kappa value was 71%, indicating that the model's level of accuracy is acceptable. Additionally, in regard to discrepancies between quantity and location, the findings revealed that the simulation models predicted the number of disease cases rather than their location, which is consistent with the results of previous studies that have used the same tool to predict LULC changes (Memarian et al. 2012). Our findings are similar to those of previous studies that reported a strong alignment between simulation results and the recorded data (Datoc et al. 2016; Enduri and Jolad 2018; Eosina et al. 2016; Medeiros et al. 2011; Ortigoza et al. 2019; Pereira and Schimit 2018; Santos et al. 2009; Sun et al. 2010; Theodorakos et al. 2017).

The proposed approach demonstrated several strengths. First, the methodological framework could be applied to other vector-borne diseases in other regions with different spatial scales and different factors that influence transmit of the disease under investigation. Second, although the proposed model does not explain the occurrence of the disease in terms of the spatial and temporal patterns, this study confirmed that the prediction of future cases is possible because of the high accuracy of the simulated risk maps when compared with recorded data. Moreover, health workers or epidemiologists can obtain information about spatial risk levels by taking into account the temporal characteristics when generating hypotheses for future research.

Despite the overall accuracy of the simulated maps using the proposed tool to predict DF hotspot locations, they have not been used sufficiently for the planning of

dengue control intervention. In order for these models to be used appropriately, it is important to understand their underlying assumptions. The MOLUSCE plugin has a simple interface which allows the user to create numerous scenarios to obtain a better understanding of the potential influence of significant factors on DF disease. The ability to change the input values allows the simulation of various scenarios of dengue transmission patterns, which gives a better understanding of the role of certain factors during the onset of epidemics. This understanding can help decision-makers to determine the steps that should be taken to avoid the transmission of epidemics. It is anticipated that this study will provide a foundation for future research that could model different aspects of DF.

4.6 Summary

This chapter presents and discusses the results obtained by the proposed framework to predict DF cases from a spatiotemporal perspective. Jeddah city was used as the case study, district boundaries as the spatial scale and annual reported confirmed cases as the temporal scale. The proposed framework led to the following results:

- The proposed models explicitly retrieved and illustrated the main aspects that influence DF illness to be employed in DF Spatiotemporal prediction models. In addition, the evaluation of the city of Jeddah in Saudi Arabia was offered. Different used software and their applications in this study were explained, as well as the datasets and their origins.
- The impact of the data quality and missing values in the prediction model performance was demonstrated and assessed. Moreover, several analyses and algorithms using historical dengue cases were applied to assess hot spot districts and perform spatiotemporal prediction models.

- In regard to the first objective, the common factors influencing DF disease were identified in the literature and used for spatiotemporal analysis to provide a better understanding of confirmed DF cases in the presence of MD for Jeddah city. In addition, clusters were created of patients with comparable characteristics by adopting an advanced machine-learning approach. Moreover, hotspot districts in the study area were obtained through several spatial analyses.
- To achieve the second objective, the accuracy of the DF prediction model was assessed by investigating the quality of the input data. A better accuracy performance was achieved, three scenarios and several machine learning prediction approaches were compared. The results demonstrate the efficiency of the cluster-based technique in increasing the accuracy of the prediction model.
- In regard to the last objective, a simulation of disease future threats and risk districts was performed using the cellular automata approach for the study area. The adopted approach provides risk maps with an accuracy of over 92% based on the recorded confirmed cases.
- The findings observed in the proposed models in terms of significant factors, accuracy, limitations, and strength were discussed.

The next chapter concludes the thesis with a discussion of the research contributions, the shortcomings and limitations of this study, and recommendations for future research directions.

CHAPTER 5

CONCLUSIONS AND FUTURE WORK RECOMMENDATIONS

5.1 Introduction

The current thesis deployed a comprehensive framework for spatiotemporal modelling of dengue fever (DF) to achieve a high-performance prediction model. Each modelling stage was investigated and analysed, from the initial data collection to the testing of the final model performance. An essential step in the spatiotemporal prediction model is to unify the variables' scale. Taking Jeddah city as a case study, this work simulated DF spatiotemporal risk maps based on confirmed cases of the disease using one year and the district's boundaries as the temporal and spatial scales respectively. The methodology included a novel combination of machine-learning algorithms and GIS-based techniques to improve the accuracy of a DF spatiotemporal prediction model. DF is the second-highest life-threatening infection disease globally and, to date, there is no vaccine for it. Thus, effective and efficient surveillance systems are essential for controlling dengue virus disease using appropriate spatial and temporal scales. Moreover, a combination of GIS and remote sensing technologies, together with advanced and statistical analysis approaches, may improve the DF prediction systems. Despite previous works to predict DF spatiotemporal patterns, failure to develop a comprehensive prediction framework is considered a significant research gap. The novel developed framework and the adopted methods here could be used for other vector-bone disease (VBD) diseases and different spatiotemporal scales. The results of this study may assist health authorities in developing better preventive strategies and increasing public interventions' effectiveness.

5.2 Objective 1 (Develop a data analytical model in the presence of MD): conclusions

In this study, a robust data analysis model was created to provide a better understanding of confirmed DF cases despite missing data (MD), and to obtain better insights into risk factors associated with confirmed cases. Moreover, by means of machine learning (ML), clusters of patients with comparable characteristics were created. This was accomplished with a self-organizing feature map (SOFM) and the density-based spatial clustering of applications with noise (DBSCAN). This study used remote sensing (RS) and geographical information system (GIS) technologies to gather important data about the spatial and temporal patterns of DF, and to determine the impact of associated risk factors on the prediction of DF cases in Jeddah city. The statistical analysis of the case averages for the period being investigated shows that the number of cases begins to increase from March until May, which has the highest number of cases. The number of cases in May is greater than other months followed by June, after which the numbers decrease until they reach the lowest level in October. In addition, spatial analysis was conducted using the ordinary least square (OLS) and geographically weighted regression (GWR) models to identify high-risk areas on an annual basis. The finding shows that there is a significant concentration of the disease in the central districts of the city according to the ordinary least squares (OLS). Overall, OLS outperforms GWR when identifying hotspot areas based on the collected data and performed analysis.

This study contributes to this area of research by applying successful methods to deal with MD, which improves the data quality without producing data bias or negatively affecting the accuracy of the prediction model. To increase the efficiency of MD imputation, SOFM was adopted due to its efficiency in dealing with MD and reducing the data to two dimensions. Then DBSCAN was applied to divide the new data obtained by SOFM into different clusters. The MD for categorical features were replaced using the

most frequent values in that cluster, where the highest mean values were used to replace missing numerical values. Several methods were adopted to model each cluster and determine the greatest accuracy. The methods applied to classify confirmed cases were: Decision Tree, k-nearest neighbours, Random Forest, AdaBoost, Support Vector Classification (SVC), CatBoost, and Naive Bayes. The findings demonstrated that the CatBoost classifier achieved the best accuracy for analysing confirmed cases. SOFM reduced the data into low-dimension, and then the DBSCAN algorithm group patients with similar features into clusters; DBSCAN detected and retrieved six clusters from this data. The clustering of confirmed cases increases CatBoost's modelling prediction accuracy and reveals complex factors that influence this accuracy. Because confirmed cases in each cluster have different features, CatBoost is applied to each cluster individually to improve the prediction accuracy. Variable values in each cluster are analysed to clarify the confirmed cases of a specific subset of DF incidents. Moreover, the modelling of each group of patient clusters could improve the modelling accuracy under particular circumstances. It can be concluded that the CatBoost model achieved a good level of accuracy in predicting DF cases. The proposed novel, data-driven and machine-learning-based strategy facilitates the understanding and identification of patterns associated with confirmed DF cases. The study's findings can be utilized to cluster historical patient data into groups or subgroups consisting of similar variables. Using identifiable patient clusters rather than raw historical data improves the model accuracy provided by CatBoost. In short, to the best of our knowledge, this is the first study to apply dengue prediction modelling to investigate the impact of missing values in the models, and to propose a comprehensive framework to improve the prediction accuracy.

5.3 Objective 2 (Improve MD imputation and prediction model performance): conclusions

In this study, a novel approach is proposed for the imputation of DF spatiotemporal data. A cluster-based technique was applied to determine the appropriate values that should replace any missing values. The accuracy of the DF model depends on the quality of the input data, which may contain errors caused by inconsistencies, misinformation or lack of information due to MD. Thus, the current study analysed the impact of missing values, and several methods were applied to handle this issue. Moreover, the study developed a novel imputation method using ML techniques in order to obtain more accurate MD imputations. The proposed approach applied SOFM and DBSCAN clustering methods to produce clusters of dengue cases with similar features in order to increase the accuracy of the missing values imputed in each cluster, thereby improving the overall prediction accuracy of the model. The proposed methods were compared to traditional imputation methods by assessing the performance of several prediction models including Decision Tree, K-Nearest Neighbors (KNN), Random Forest, AdaBoost, Support Vector Classification (SVC), CatBoost, and Naive Bayes using different experimental scenarios. The AdaBoost and CatBoost models obtained high accuracy for cluster 8 (2014), while the KNN model also scored 100% accuracy for cluster 7 (2017). This study makes a significant contribution by improving prediction accuracy. Utilizing the results obtained from several scenarios, the performance of the proposed approach was compared with those of several traditional and complex imputation algorithms. The study findings show that, based on the modelling of each cluster separately for each year, the accuracy of the prediction model was improved to achieve a high prediction accuracy, demonstrating that the proposed imputation method outperforms traditional methods. However, the choice of algorithms (i.e., traditional or ML-based) should depend on the

reason for addressing the problem of MD. Lastly, this study addressed a gap in the research by demonstrating that the separate modelling for each cluster generated by SOFM-DBSCAN can improve prediction accuracy and offer a better understanding of the factors associated with the disease in each cluster.

5.4 Objective 3 (Simulate risk areas): conclusions

In this study, the main objectives were to examine the spatiotemporal aspects of DF cases at the district scale, and to validate the efficiency of the MOLUSCE-plugin in predicting future threats and identifying hotspot districts. Several scenarios were simulated in order to achieve better prediction performance. For decades, the convergence of spatial statistics and simulation has driven a great deal of the research on geographic phenomena. However, the temporal aspects of geographic data have been given little attention in the domain of geographic information science and simulation in Jeddah, Saudi Arabia. To obtain better insights regarding the prediction of DF cases, this study focused on the analysis of spatiotemporal patterns and proposed innovative ways to present the results for the location and time dimensions associated with Jeddah city. Additionally, this study examined the feasibility of using cellular automata through the QGIS MOLUSCE plugin to investigate spatiotemporal issues. Cellular automata were applied here to achieve the desired objectives using the MOLUSCE plugin with QGIS software. Furthermore, the 2018 map of the region of interest containing recorded cases, as well as the Kappa coefficient, were utilised to determine the accuracy of the model responses. The tool's ability to predict future events was tested via three proposed scenarios, and the findings were compared with the observations contained in the 2018 map. The results of the simulations indicate that the tool can predict future threats with an accuracy just above 92%. However, the prediction map generated with the MOLUSCE

plugin is able to predict the number of confirmed DF cases more correctly than it does the location of these cases. Results from the analysis of multi-temporal dengue cases over a seven-year period show that population distribution is significantly correlated with the number of predicted dengue cases for a particular year. The findings indicate that the proposed model is capable of predicting potential DF risks with a high level of accuracy. Results suggest that simulation is a valuable means of obtaining a better understanding of space-time complexity in dynamic spatial phenomena, such as DF. The results presented in this objective make two important contributions. First, the MOLUSCE tool was adjusted to fit the related spatiotemporal fields, and the efficiency of the model in predicting future phenomena other than land changes, was confirmed. Second, the efficiency of the tool was validated as a means of predicting dengue risk areas using historical data.

5.5 Contributions

Previous studies have attempted to analyse DF patterns by means of spatiotemporal modelling. However, none established a comprehensive framework that involves all the modelling stages starting from data collection stage to the final model performance and visualization. This is a significant gap in the previous research. Moreover, although information can be obtained from several sources, it is difficult to find data that is both accurate and complete. Also, inefficient data pre-processing can result in inadequate modelling due to bias or inaccuracy. This makes it challenging to produce a model with optimal performance that will be acceptable to decision-makers and stakeholders. Hence, in this thesis, a modelling methodology is proposed to obtain detailed and accurate spatiotemporal patterns that, together with a GIS-based explanation, will validate the efficiency of the proposed framework. In this current research, a novel framework is

designed that combines ML and GIS to develop a high-performance model for the prediction of DF based on spatiotemporal factors pertaining to a district rather than a larger geographic area. The thesis's main contributions are:

- This thesis is the first attempt to propose a comprehensive framework for the spatiotemporal modelling of DF to achieve a high-performance prediction by investigating and analysing each modelling stage from the initial collection of the data to the testing of the model's performance. Moreover, this analysis is also the first to explicitly consider the impact of the DF data collection process and data quality on DF spatiotemporal models.
- Unlike previous modelling studies, these results demonstrate in detail the importance of each stage of modelling, and its impact on the final model. This is done by linking the quality of the collected data and its impact on confirmed disease cases, taking a district's boundaries as the spatial scale and the total annual reported cases as the temporal scale. This thesis contributes to the current body of knowledge by applying successful methods to deal with MD, which improves the data quality without producing data bias or negatively affecting the accuracy of the prediction model.
- Then, to increase the efficiency of MD imputation, both simple and advanced imputation methods are applied and the results are compared to determine the best algorithm. Moreover, the study findings show that, based on the modelling of each cluster separately for each year, adopted methods were able to improve the accuracy of the prediction model to achieve a high prediction accuracy, demonstrating that the proposed imputation method outperforms traditional methods. In addition, the findings demonstrate that the model can improve prediction accuracy under different scenarios and circumstances.

- For prediction modelling, cellular automata using MOLUSCE, a QGIS plugin was applied to evaluate DF confirmed cases changes over time. This enabled the assessment of the efficiency of the data preparation and the determination of the likelihood of future threats of DF. Findings indicate that the used approach, based on the historical data, produces prediction maps that have over 92% accuracy in terms of forecasting future threats. By way of summary, it is anticipated that the results presented in this thesis will be disseminated in reputable journals and can be generalised to other modelling studies.

5.6 Research limitations

The proposed framework for modelling DF disease has been applied using Jeddah city as the study area and achieved the research objectives. Despite the thesis contribution and finding, there are several limitations to achieve each of the objectives outlined and discussed below:

- The proposed framework was created to investigate the confirmed and associated risk factors for DF in the presence of missing values. Although the study achieved the first and second objectives, it has several limitations.
  1. The demographic data contains information such as ‘occupation’ for some years (2012 and 2013) but not for the remaining years of the study period.
  2. The main significant issue in the demographic data is the absence of an accurate address or district name, which necessitated a long process to adjust these data accordingly. Therefore, due to the lack of some demographic factors, they could not be included in the work related to the first objective.
  3. Regarding climatic features, only two weather stations cover Jeddah city. The problem with the data obtained from the General Authority of Meteorology

and Environmental is that it covers a city, not a district. In other words, it returns the same result for all climatic features for all the districts, rendering these data useless. Therefore, annual data from satellite images taken during the period of interest were collected.

4. Vector data and host "human" data were unavailable for the study areas, and such data is required for more accurate prediction.
5. The limited amount of data used in the study for modelling DF cases using ML algorithms is another limitation of current work.

- Several limitations in terms of the second objective warrant discussion.

1. ML methods require a significant amount of time for imputations, as does the gathering of a comprehensive dataset for model training under several scenarios designed for the experiments.
2. Determining the optimal parameters for algorithms/models is challenging for a particular dataset. Moreover, the hyperparameters chosen for model training affect the model's accuracy (Claesen and De Moor 2015). Therefore, the current study applied default parameters in all algorithms and compared the models' accuracy using those values, and used a hyperparameter that contains a series of other values for each parameter.

- Regarding the third objective, two shortcomings should be acknowledged and addressed in future research.

1. In order to better interpret spatiotemporal variations in the DF epidemic, other influential factors should be explored and included in the simulation model. The CA lattice and state transition rules in CA models naturally account for the

spatial features of host and vector populations, including mobility patterns and heterogeneities (Dias and Monteiro 2018).

2. Although the adopted tool shows good performance in simulating the risk map for 2018 using historical observed data in numerous scenarios, the CA model does not always directly predict the level of risk (e.g., high/medium etc.) in a certain area in terms of the spatial dimension. In other words, although the proposed model confirmed that the prediction of future cases is possible given the high percentage of correct mapping based on recorded data, this study does not explain the occurrence of the disease in terms of its spatial and temporal patterns.

5.7 Recommendation for future work

Despite the aforementioned limitations and drawbacks, all the proposed models in this thesis achieved the desired objectives. The thesis' overall goal was to develop a comprehensive framework and estimate with a high level of accuracy any future risk on a district scale in order to better implement and control disease-containment procedures. However, further research can be conducted regarding DF spatiotemporal prediction models. The research and recommendations for future work are:

- It is recommended that future researchers collect comprehensive climatic, epidemiological, environmental, and demographic information from local authorities and compare it with data obtained by satellite images.
- In order to obtain an optimal prediction method, several prediction models could be designed and evaluated. Moreover, previous works have demonstrated that deep learning models outperform traditional machine learning classifiers

(Onan 2020). Therefore, future work should consider the application of deep learning-based models to improve performance.

- Features selection is a major aspect of machine learning which requires determining the variables that will fit the desired classification model (Onan 2015). Hence, additional feature selection algorithms need further investigation.
- In order to obtain a more precise result from cluster-based analysis, it is recommended that future research investigate the impact of cluster size on the model's performance.
- In order to derive comprehensive information, the proposed methodological framework could be applied to other VBDs in other regions using different spatial scales and different factors that could influence the transmission of the disease.
- Although the overall accuracy of the prediction map using CA was above 92%, future research could consider a different temporal scale of dengue patterns and formulate hypotheses to investigate this aspect in more detail.

REFERENCES

Abou El-Saoud, W., Gabr, S.S., Abdel-Rahim, I.H. & Morsy, E. 2018, 'Determination of vectors' reproduction spots in Makkah using GIS and remote sensing techniques', *The Scientific Bulletin*, p. 43.

Acharya, B.K., Cao, C., Lakes, T., Chen, W. & Naeem, S. 2016, 'Spatiotemporal analysis of dengue fever in Nepal from 2010 to 2014', *BMC Public Health*, vol. 16, no. 1, pp. 1-10.

Acharya, B.K., Cao, C., Lakes, T., Chen, W., Naeem, S. & Pandit, S. 2018, 'Modeling the spatially varying risk factors of dengue fever in Jhapa district, Nepal, using the semi-parametric geographically weighted regression model', *International Journal of Biometeorology*, vol. 62, no. 11, pp. 1973-86.

Ahmed, A. & Hannan, S.A. 2012, 'Data mining techniques to find out heart diseases: an overview', *International Journal of Innovative Technology and Exploring Engineering (IJITEE)*, vol. 1, no. 4, pp. 18-23.

Ahmed, K.N. & Razak, T.A. 2014, 'A comparative study of different density based spatial clustering algorithms', *International Journal of Computer Applications*, vol. 975, p. 8887.

Ahmed, Q.A., Arabi, Y.M. & Memish, Z.A. 2006, 'Health risks at the Hajj', *The Lancet*, vol. 367, no. 9515, pp. 1008-15.

Akter, R., Hu, W., Gatton, M., Bambrick, H., Cheng, J. & Tong, S. 2021, 'Climate variability, socio-ecological factors and dengue transmission in tropical Queensland, Australia: a Bayesian spatial analysis', *Environmental Research*, vol. 195, p. 110285.

Akter, R., Naish, S., Gatton, M., Bambrick, H., Hu, W. & Tong, S. 2019, 'Spatial and temporal analysis of dengue infections in Queensland, Australia: recent trend and perspectives', *PLOS ONE*, vol. 14, no. 7, p. e0220134.

Akter, R., Naish, S., Hu, W. & Tong, S. 2017, 'Socio-demographic, ecological factors and dengue infection trends in Australia', *PLOS ONE*, vol. 12, no. 10, p. e0185551.

Al-Azraqi, T.A., El Mekki, A.A. & Mahfouz, A.A. 2013, 'Seroprevalence of dengue virus infection in Aseer and Jizan regions, southwestern Saudi Arabia', *Transactions of the Royal Society of Tropical Medicine and Hygiene*, vol. 107, no. 6, pp. 368-71.

Al-Hagery, M.A., Alreshoodi, L.A., Almutairi, M.A., Alsharekh, S.I. & Alkhowaiter, E.S. 2019, 'A hybrid technique for cleaning missing and misspelling Arabic data in data warehouse', *International Journal of Information Technology and Computer Science*, pp. 17-25.

Al-Raddadi, R., Alwafi, O., Shabouni, O., Akbar, N., Alkhalawi, M., Ibrahim, A., Hussain, R., Alzahrani, M., Al Helal, M. & Assiri, A. 2019, 'Seroprevalence of dengue fever and the associated sociodemographic, clinical, and environmental factors in Makkah, Madinah, Jeddah, and Jizan, Kingdom of Saudi Arabia', *Acta Tropica*, vol. 189, pp. 54-64.

Al-Sarem, M., Saeed, F., Boulila, W., Emara, A.H., Al-Mohaimeed, M. & Errais, M. 2021, 'Feature selection and classification using CatBoost method for improving the performance of predicting Parkinson's disease', *Advances on Smart and Soft Computing*, Springer, pp. 189-99.

Al-Tayib, O.A. 2019, 'An overview of the most significant zoonotic viral pathogens transmitted from animal to human in Saudi Arabia', *Pathogens*, vol. 8, no. 1, p. 25.

Al Kafy, A., Al Rakib, A., Akter, K.S., Rahaman, Z.A., Jahir, D.M., Subramanyam, G., Michel, O.O. & Bhatt, A. 2021, 'The operational role of remote sensing in assessing and predicting land use/land cover and seasonal land surface temperature using machine learning algorithms in Rajshahi, Bangladesh', *Applied Geomatics*, vol. 13, no. 4, pp. 793-816.

Al Masud, S.M.R., Bakar, A.A. & Yussof, S. 2018, 'A systematic review of technological issues in monitoring pilgrims' health during hajj: current state, challenges and future directions', *Journal of Theoretical & Applied Information Technology*, vol. 96, no. 7.

Aleeban, M. & Mackey, T.K. 2016, 'Global health and visa policy reform to address dangers of Hajj during summer seasons', *Frontiers in Public Health*, vol. 4, p. 280.

Alhaeli, A., Bahkali, S., Ali, A., Househ, M.S. & El-Metwally, A.A. 2016, 'The epidemiology of dengue fever in Saudi Arabia: a systematic review', *Journal of Infection and Public Health*, vol. 9, no. 2, pp. 117-24.

Aljoufie, M. & Tiwari, A. 2021, 'Modeling road safety in car-dependent cities: case of Jeddah city, Saudi Arabia', *Sustainability*, vol. 13, no. 4, p. 1816.

Alkhaldy, I. 2017, 'Modelling the association of dengue fever cases with temperature and relative humidity in Jeddah, Saudi Arabia—A generalised linear model with break-point analysis', *Acta Tropica*, vol. 168, pp. 9-15.

Alshehri, A. 2019, *A machine learning approach to predicting community engagement on social media during disasters*, University of South Florida.

Altassan, K.K., Morin, C., Shocket, M.S., Ebi, K. & Hess, J. 2019, 'Dengue fever in Saudi Arabia: a review of environmental and population factors impacting emergence and spread', *Travel Medicine and Infectious Disease*, vol. 30, pp. 46-53.

Althouse, B.M., Ng, Y.Y. & Cummings, D.A. 2011, 'Prediction of dengue incidence using search query surveillance', *PLOS Neglected Tropical Diseases*, vol. 5, no. 8, p. e1258.

Alwafi, O.M. 2013, 'Dengue fever in Makkah, Kingdom of Saudi Arabia, 2008-2012', *American Journal of Research Communication*, pp. 123-39.

Alzubi, J., Nayyar, A. & Kumar, A. 2018, 'Machine Learning from Theory to Algorithms: An Overview', *Journal of Physics: Conference Series*, vol. 1142, p. 012012.

Andraud, M., Hens, N., Marais, C. & Beutels, P. 2012, 'Dynamic epidemiological models for dengue transmission: a systematic review of structural approaches', *PLOS ONE*, vol. 7, no. 11, p. e49085.

Andre, F.E., Booy, R., Bock, H.L., Clemens, J., Datta, S.K., John, T.J., Lee, B.W., Lolekha, S., Peltola, H. & Ruff, T. 2008, 'Vaccination greatly reduces disease, disability, death and inequity worldwide', *Bulletin of the World Health Organization*, vol. 86, pp. 140-6.

Aneesha Satya, B., Shashi, M. & Deva, P. 2020, 'Future land use land cover scenario simulation using open source GIS for the city of Warangal, Telangana, India', *Applied Geomatics*, vol. 12, no. 3, pp. 281-90.

Arboleda, S., Jaramillo-O, N. & Peterson, A.T. 2012, 'Spatial and temporal dynamics of Aedes aegypti larval sites in Bello, Colombia', *Journal of Vector Ecology*, vol. 37, no. 1, pp. 37-48.

Asan, U. & Ercan, S. 2012, 'An introduction to self-organizing maps', *Computational Intelligence Systems in Industrial Engineering*, Springer, pp. 295-315.

Ashby, J., Moreno-Madriñán, M.J., Yiannoutsos, C.T. & Stanforth, A. 2017, 'Niche modeling of dengue fever using remotely sensed environmental factors and boosted regression trees', *Remote Sensing*, vol. 9, no. 4, p. 328.

Ashshi, A.M. 2015, 'Serodetection of dengue virus and its antibodies among blood donors in the western region of Saudi Arabia: a preliminary study', *Blood Transfusion*, vol. 13, no. 1, p. 135.

Ashshi, A.M. 2017, 'The prevalence of dengue virus serotypes in asymptomatic blood donors reveals the emergence of serotype 4 in Saudi Arabia', *Virology Journal*, vol. 14, no. 1, pp. 1-8.

Astuti, E.P., Dhewantara, P.W., Prasetyowati, H., Ipa, M., Herawati, C. & Hendrayana, K. 2019, 'Paediatric dengue infection in Cirebon, Indonesia: a temporal and spatial analysis of notified dengue incidence to inform surveillance', *Parasites & Vectors*, vol. 12, no. 1, pp. 1-12.

Aswi, A., Cramb, S., Moraga, P. & Mengersen, K. 2019, 'Bayesian spatial and spatio-temporal approaches to modelling dengue fever: a systematic review', *Epidemiology & Infection*, vol. 147.

Attaway, D.F., Jacobsen, K.H., Falconer, A., Manca, G. & Waters, N.M. 2016, 'Risk analysis for dengue suitability in Africa using the ArcGIS predictive analysis tools (PA tools)', *Acta Tropica*, vol. 158, pp. 248-57.

Ayyub, M., Khazindar, A.M., Lubbad, E.H., Barlas, S., Alfi, A.Y. & Al-Ukayli, S. 2006, 'Characteristics of dengue fever in a large public hospital, Jeddah, Saudi Arabia', *Journal of Ayub Medical College Abbottabad*, vol. 18, no. 2, pp. 9-13.

Aziz, A.T., Al-Shami, S.A., Mahyoub, J.A., Hatabbi, M., Ahmad, A.H. & Rawi, C.S.M. 2014, 'An update on the incidence of dengue gaining strength in Saudi Arabia and current control approaches for its vector mosquito', *Parasites & Vectors*, vol. 7, no. 1, pp. 1-4.

Badreddine, S., Al-Dhaheri, F., Al-Dabbagh, A., Al-Amoudi, A., Al-Ammari, M., Elatassi, N., Abbas, H., Magliah, R., Malibari, A. & Almoallim, H. 2017, 'Dengue fever: clinical features of 567 consecutive patients admitted to a tertiary care center in Saudi Arabia', *Saudi Medical Journal*, vol. 38, no. 10, p. 1025.

Bai, L., Morton, L.C. & Liu, Q. 2013, 'Climate change and mosquito-borne diseases in China: a review', *Globalization and Health*, vol. 9, no. 1, pp. 1-22.

Baidya, D., Sarkar, A., Mondal, A. & Mitra, D. 2021, 'Application of Cellular Automata (CA) for Predicting Urban Growth and Disappearance of Vegetation and Waterbodies', paper presented to the *Proceedings of Research and Applications in Artificial Intelligence*, Singapore.

Baker-Yeboah, S. & Kilpatrick, K.A. 2016, 'Pathfinder Version 5.3 AVHRR Sea Surface Temperature Climate Data Record', paper presented to the *AGU Fall Meeting Abstracts*.

Baraldi, A.N. & Enders, C.K. 2010, 'An introduction to modern missing data analyses', *Journal of School Psychology*, vol. 48, no. 1, pp. 5-37.

Batista, G.E. & Monard, M.C. 2002, 'A study of K-nearest neighbour as an imputation method', *His*, vol. 87, no. 251-260, p. 48.

Batista, G.E. & Monard, M.C. 2003, 'An analysis of four missing data treatment methods for supervised learning', *Applied Artificial Intelligence*, vol. 17, no. 5-6, pp. 519-33.

Belete, D.M. & Huchaiah, M.D. 2021, 'Grid search in hyperparameter optimization of machine learning models for prediction of HIV/AIDS test results', *International Journal of Computers and Applications*, pp. 1-12.

Bergstra, J. & Bengio, Y. 2012, 'Random search for hyper-parameter optimization', *Journal of Machine Learning Research*, vol. 13, no. 2.

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. 2013, 'The global distribution and burden of dengue', *Nature*, vol. 496, no. 7446, pp. 504-7.

Biamonte, J., Wittek, P., Pancotti, N., Rebentrost, P., Wiebe, N. & Lloyd, S. 2017, 'Quantum machine learning', *Nature*, vol. 549, no. 7671, pp. 195-202.

Birant, D. & Kut, A. 2007, 'ST-DBSCAN: an algorithm for clustering spatial–temporal data', *Data & Knowledge Engineering*, vol. 60, no. 1, pp. 208-21.

Bouzid, M., Colón-González, F.J., Lung, T., Lake, I.R. & Hunter, P.R. 2014, 'Climate change and the emergence of vector-borne diseases in Europe: case study of dengue fever', *BMC Public Health*, vol. 14, no. 1, pp. 1-12.

Bowman, L.R., Runge-Ranzinger, S. & McCall, P. 2014, 'Assessing the relationship between vector indices and dengue transmission: a systematic review of the evidence', *PLOS Neglected Tropical Diseases*, vol. 8, no. 5, p. e2848.

Bright, E., Rose, A. & Urban, M. 2013, 'LandScan Global 2012', electronic data set, <landscan.ornl.gov>.

Bright, E., Rose, A. & Urban, M. 2014, 'LandScan Global 2013', electronic data set, <landscan.ornl.gov>.

Bright, E., Rose, A. & Urban, M. 2015, 'LandScan Global 2014', electronic data set, <landscan.ornl.gov>.

Bright, E., Rose, A. & Urban, M. 2016, 'LandScan Global 2015', electronic data set, <landscan.ornl.gov>.

Bright, E., Rose, A., Urban, M. & McKee, J. 2017, 'LandScan Global 2016', electronic data set, <landscan.ornl.gov>.

Brown, M.L. & Kros, J.F. 2003, 'Data mining and the impact of missing data', *Industrial Management & Data Systems*.

Buczak, A.L., Baugher, B., Babin, S.M., Ramac-Thomas, L.C., Guven, E., Elbert, Y., Koshute, P.T., Velasco, J.M.S., Roque Jr, V.G. & Tayag, E.A. 2014, 'Prediction of high incidence of dengue in the Philippines', *PLOS Neglected Tropical Diseases*, vol. 8, no. 4, p. e2771.

Buczak, A.L., Koshute, P.T., Babin, S.M., Feighner, B.H. & Lewis, S.H. 2012, 'A data-driven epidemiological prediction method for dengue outbreaks using local and remote sensing data', *BMC Medical Informatics and Decision Making*, vol. 12, no. 1, pp. 1-20.

Cao, Z., Liu, T., Li, X., Wang, J., Lin, H., Chen, L., Wu, Z. & Ma, W. 2017, 'Individual and interactive effects of socio-ecological factors on dengue fever at fine spatial scale: a geographical detector-based analysis', *International Journal of Environmental Research and Public Health*, vol. 14, no. 7, p. 795.

Carbajo, A.E., Cardo, M.V. & Vezzani, D. 2012, 'Is temperature the main cause of dengue rise in non-endemic countries? The case of Argentina', *International Journal of Health Geographics*, vol. 11, no. 1, pp. 1-11.

Cauthen, K.R., Lambert, G., Ray, J. & Lefantzi, S. 2016, *Imputing data that are missing at high rates using a boosting algorithm*, Sandia National Lab.(SNL-NM), Albuquerque, NM (United States); Sandia National Lab.(SNL-CA), Livermore, CA (United States).

Chapelle, O., Vapnik, V., Bousquet, O. & Mukherjee, S. 2002, 'Choosing multiple parameters for support vector machines', *Machine Learning*, vol. 46, no. 1, pp. 131-59.

Chen, S.-C. & Hsieh, M.-H. 2012, 'Modeling the transmission dynamics of dengue fever: implications of temperature effects', *Science of the Total Environment*, vol. 431, pp. 385-91.

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 Medicine*, vol. 16, no. 1, pp. 1-13.

Cheong, Y.L., Burkart, K., Leitão, P.J. & Lakes, T. 2013, 'Assessing weather effects on dengue disease in Malaysia', *International Journal of Environmental Research and Public Health*, vol. 10, no. 12, pp. 6319-34.

Claesen, M. & De Moor, B. 2015, 'Hyperparameter search in machine learning', *arXiv preprint arXiv:1502.02127*.

Cucunawangsih & Lugito, N.P.H. 2017, 'Trends of dengue disease epidemiology', *Virology: Research and Treatment*, vol. 8, p. 1178122X17695836.

Cui, H. & Bai, J. 2019, 'A new hyperparameters optimization method for convolutional neural networks', *Pattern Recognition Letters*, vol. 125, pp. 828-34.

Cummings, D.A., Irizarry, R.A., Huang, N.E., Endy, T.P., Nisalak, A., Ungchusak, K. & Burke, D.S. 2004, 'Travelling waves in the occurrence of dengue haemorrhagic fever in Thailand', *Nature*, vol. 427, no. 6972, pp. 344-7.

Dasgupta, S., Sharma, N., Sinha, S. & Raghavendra, S. 2019, 'Evaluating the performance of machine learning using feature selection methods in a dengue dataset', *International Journal of Engineering and Advanced Technology (IJEAT)*, vol. 8, no. 5, pp. 2679-85.

Datoc, H.I., Caparas, R. & Caro, J. 2016, 'Forecasting and data visualization of dengue spread in the Philippine Visayas Island group', paper presented to the *In 2016 7th International Conference on Information, Intelligence, Systems & Applications (IISA) (pp. 1-4). IEEE*.

De Lima, T.F.M., Lana, R.M., de Senna Carneiro, T.G., Codeço, C.T., Machado, G.S., Ferreira, L.S., de Castro Medeiros, L.C. & Davis Junior, C.A. 2016, 'Dengueme: a tool for the modeling and simulation of dengue spatiotemporal dynamics', *International Journal of Environmental Research and Public Health*, vol. 13, no. 9, p. 920.

Defourny, P., Brockmann, C., Bontemps, S., Lamarche, C., Santoro, M., Boettcher, M. & Wevers, J. 2017, *CCI-LC PUGv2 Phase II. Land cover climate change initiative-product user guide v2*, Technical report.

Delmelle, E., Hagenlocher, M., Kienberger, S. & Casas, I. 2016, 'A spatial model of socioeconomic and environmental determinants of dengue fever in Cali, Colombia', *Acta Tropica*, vol. 164, pp. 169-76.

Deng, Y., Wilson, J.P. & Bauer, B. 2007, 'DEM resolution dependencies of terrain attributes across a landscape', *International Journal of Geographical Information Science*, vol. 21, no. 2, pp. 187-213.

Depradine, C. & Lovell, E. 2004, 'Climatological variables and the incidence of dengue fever in Barbados', *International Journal of Environmental Health Research*, vol. 14, no. 6, pp. 429-41.

Dhewantara, P.W., Marina, R., Puspita, T., Ariati, Y., Purwanto, E., Hananto, M., Hu, W. & Magalhaes, R.J.S. 2019, 'Spatial and temporal variation of dengue incidence in the island of Bali, Indonesia: an ecological study', *Travel Medicine and Infectious Disease*, vol. 32, p. 101437.

Dhimal, M., Ahrens, B. & Kuch, U. 2015, 'Climate change and spatiotemporal distributions of vector-borne diseases in Nepal–a systematic synthesis of literature', *PLOS ONE*, vol. 10, no. 6, p. e0129869.

Dias, J. & Monteiro, L.H. 2018, 'Clustered breeding sites: shelters for vector-borne diseases', *Computational and Mathematical Methods in Medicine*, vol. 2018.

Dieng, H., Ahmad, A.H., Mahyoub, J.A., Turkistani, A.M., Mesed, H., Koshike, S., Satho, T., Salmah, M.C., Ahmad, H. & Zuharah, W.F. 2012, 'Household survey of container–breeding mosquitoes and climatic factors influencing the prevalence of Aedes aegypti (Diptera: Culicidae) in Makkah city, Saudi Arabia', *Asian Pacific Journal of Tropical Biomedicine*, vol. 2, no. 11, pp. 849-57.

Domingo, C., Niedrig, M., Teichmann, A., Kaiser, M., Rumer, L., Jarman, R.G. & Donoso-Mantke, O. 2010, '2nd International external quality control assessment for the molecular diagnosis of dengue infections', *PLOS Neglected Tropical Diseases*, vol. 4, no. 10, p. e833.

Dorogush, A.V., Ershov, V. & Gulin, A. 2018, 'CatBoost: gradient boosting with categorical features support', *arXiv preprint arXiv:1810.11363*.

Ducheyne, E., Tran Minh, N.N., Haddad, N., Bryssinckx, W., Buliva, E., Simard, F., Malik, M.R., Charlier, J., De Waele, V. & Mahmoud, O. 2018, 'Current and future distribution of Aedes aegypti and Aedes albopictus (Diptera: Culicidae) in WHO Eastern Mediterranean region', *International Journal of Health Geographics*, vol. 17, no. 1, pp. 1-13.

Earth Resources Observation Science Center 2018, 'Shuttle Radar Topography Mission (SRTM) 1 Arc-Second Global', electronic data set.

Eekhout, I., De Vet, H.C., Twisk, J.W., Brand, J.P., de Boer, M.R. & Heymans, M.W. 2014, 'Missing data in a multi-item instrument were best handled by multiple imputation at the item score level', *Journal of Clinical Epidemiology*, vol. 67, no. 3, pp. 335-42.

Eggensperger, K., Feurer, M., Hutter, F., Bergstra, J., Snoek, J., Hoos, H. & Leyton-Brown, K. 2013, 'Towards an empirical foundation for assessing bayesian optimization of hyperparameters', paper presented to the *NIPS Workshop on Bayesian Optimization in Theory and Practice*.

El-Badry, A.A. & Al-Ali, K.H. 2010, 'Prevalence and seasonal distribution of dengue mosquito, Aedes aegypti (Diptera: Culicidae) in Al-Madinah Al-Munawwarah, Saudi Arabia', *Journal of Entomology*, vol. 7, no. 2, pp. 80-8.

El-Badry, A.A., El-Beshbishy, H.A., Al-Ali, K.H., Al-Hejin, A.M. & El-Sayed, W.S. 2014, 'Molecular and seroprevalence of imported dengue virus infection in Al-Madinah, Saudi Arabia', *Comparative Clinical Pathology*, vol. 23, no. 4, pp. 861-8.

El-Kafrawy, S.A., Sohrab, S.S., Ela, S.A., Abd-Alla, A.M., Alhabbab, R., Farraj, S.A., Othman, N.A., Hassan, A.M., Bergoin, M. & Klitting, R. 2016, 'Multiple introductions of dengue 2 virus strains into Saudi Arabia from 1992 to 2014', *Vector-Borne and Zoonotic Diseases*, vol. 16, no. 6, pp. 391-9.

Enduri, M.K. & Jolad, S. 2018, 'Dynamics of dengue disease with human and vector mobility', *Spatial and Spatio-temporal Epidemiology*, vol. 25, pp. 57-66.

Eosina, P., Djatna, T. & Khusun, H. 2016, 'A cellular automata modeling for visualizing and predicting spreading patterns of dengue fever', *Telkomnika*, vol. 14, no. 1, p. 228.

Espinosa, M., Weinberg, D., Rotela, C.H., Polop, F., Abril, M. & Scavuzzo, C.M. 2016, 'Temporal dynamics and spatial patterns of Aedes aegypti breeding sites, in the context of a dengue control program in Tartagal (Salta province, Argentina)', *PLOS Neglected Tropical Diseases*, vol. 10, no. 5, p. e0004621.

Ester, M., Kriegel, H.-P., Sander, J. & Xu, X. 1996, 'A density-based algorithm for discovering clusters in large spatial databases with noise', paper presented to the *kdd*.

European Centre for Disease Prevention and Control 2019, *Public health risks related to communicable diseases during the hajj 2019, Saudi Arabia, 9–14 August 2019 – 2 July 2019*, ECDC, Stockholm.

Fairos, W.W., Azaki, W.W., Alias, L.M. & Wah, Y.B. 2010, 'Modelling dengue fever (DF) and dengue haemorrhagic fever (DHF) outbreak using Poisson and Negative Binomial model', *World Academy of Science, Engineering and Technology*, vol. 62.

Faisal, T., Ibrahim, F. & Taib, M.N. 2008, 'Analysis of significant factors for dengue infection prognosis using the self organizing map', paper presented to the *2008 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*.

Faisal, T., Ibrahim, F. & Taib, M.N. 2010, 'A noninvasive intelligent approach for predicting the risk in dengue patients', *Expert Systems with Applications*, vol. 37, no. 3, pp. 2175-81.

Fakeeh, M. & Zaki, A. 2001, 'Virologic and serologic surveillance for dengue fever in Jeddah, Saudi Arabia, 1994-1999', *The American Journal of Tropical Medicine and Hygiene*, vol. 65, no. 6, pp. 764-7.

Falah, N., Karimi, A. & Harandi, A.T. 2020, 'Urban growth modeling using cellular automata model and AHP (case study: Qazvin city)', *Modeling Earth Systems and Environment*, vol. 6, no. 1, pp. 235-48.

Fan, J., Lin, H., Wang, C., Bai, L., Yang, S., Chu, C., Yang, W. & Liu, Q. 2014, 'Identifying the high-risk areas and associated meteorological factors of dengue transmission in Guangdong province, China from 2005 to 2011', *Epidemiology & Infection*, vol. 142, no. 3, pp. 634-43.

Fathima, A.S. & Manimeglai, D. 2015, 'Analysis of significant factors for dengue infection prognosis using the random forest classifier', *(IJACSA) International Journal of Advanced Computer Science and Applications*, vol. 6, no. 2, pp. 240-5.

Ferdous, J. & Rahman, M.T.U. 2019, 'Geomorphological changes along coastline of Bangladesh', paper presented to the *2nd International Conference on Water and Environmental Engineering (iCWEE2019), Dhaka, Bangladesh)*.

Ferrell, A.M. & Brinkerhoff, R.J. 2018, 'Using landscape analysis to test hypotheses about drivers of tick abundance and infection prevalence with Borrelia burgdorferi', *International Journal of Environmental Research and Public Health*, vol. 15, no. 4, p. 737.

Fischer, D., Thomas, S.M., Suk, J.E., Sudre, B., Hess, A., Tjaden, N.B., Beierkuhnlein, C. & Semenza, J.C. 2013, 'Climate change effects on Chikungunya transmission in Europe: geospatial analysis of vector's climatic suitability and virus' temperature requirements', *International Journal of Health Geographics*, vol. 12, no. 1, pp. 1-12.

Flamand, C., Fabregue, M., Bringay, S., Ardillon, V., Quénel, P., Desenclos, J.-C. & Teisseire, M. 2014, 'Mining local climate data to assess spatiotemporal dengue fever epidemic patterns in French Guiana', *Journal of the American Medical Informatics Association*, vol. 21, no. e2, pp. e232-e40.

Francisco, M.E., Carvajal, T.M., Ryo, M., Nukazawa, K., Amalin, D.M. & Watanabe, K. 2021, 'Dengue disease dynamics are modulated by the combined influences of precipitation and landscape: a machine learning approach', *Science of the Total Environment*, vol. 792, p. 148406.

Freeze, J., Erraguntla, M. & Verma, A. 2018, 'Data integration and predictive analysis system for disease prophylaxis: incorporating dengue fever forecasts', paper presented to the *Proceedings of the 51st Hawaii International Conference on System Sciences*.

Friedl, M., Sulla-Menashe, Damien 2019, 'MCD12Q1 MODIS/Terra+Aqua Land Cover Type Yearly L3 Global 500m SIN Grid V006', electronic data set, <https://doi.org/10.5067/MODIS/MCD12Q1.006>.

Fuentes-Vallejo, M. 2017, 'Space and space-time distributions of dengue in a hyper-endemic urban space: the case of Girardot, Colombia', *BMC Infectious Diseases*, vol. 17, no. 1, pp. 1-16.

Funk, C., Peterson, P., Landsfeld, M., Pedreros, D., Verdin, J., Shukla, S., Husak, G., Rowland, J., Harrison, L. & Hoell, A. 2015, 'The climate hazards infrared precipitation with stations—a new environmental record for monitoring extremes', *Scientific Data*, vol. 2, no. 1, pp. 1-21.

Gamil, M.A., Eisa, Z.M., Eifan, S.A. & Al-Sum, B.A. 2014, 'Prevalence of dengue fever in Jizan area, Saudi Arabia', *Journal of Pure and Applied Microbiology*, vol. 8, no. 1, pp. 225-31.

Gao, P., Pilot, E., Rehbock, C., Gontariuk, M., Doreleijers, S., Wang, L., Krafft, T., Martens, P. & Liu, Q. 2021, 'Land use and land cover change and its impacts on dengue dynamics in China: a systematic review', *PLOS Neglected Tropical Diseases*, vol. 15, no. 10, p. e0009879.

García-Laencina, P.J., Sancho-Gómez, J.-L. & Figueiras-Vidal, A.R. 2010, 'Pattern classification with missing data: a review', *Neural Computing and Applications*, vol. 19, no. 2, pp. 263-82.

General Authority for Statistics 2018, *The total number of pilgrims in 1439H*, GASTAT, viewed 13/05/2020, < https://www.stats.gov.sa/en/news/280>.

Germano, T. 1999, *Self organizing maps*, <Available in http://davis.wpi.edu/matt/courses/soms>.

Gibbons, R.V. & Vaughn, D.W. 2002, 'Dengue: an escalating problem', *BMJ*, vol. 324, no. 7353, pp. 1563-6.

Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D. & Moore, R. 2017, 'Google Earth Engine: Planetary-scale geospatial analysis for everyone', *Remote Sensing of Environment*, vol. 202, pp. 18-27.

Gubler, D.J. 1998, 'Dengue and dengue hemorrhagic fever', *Clinical Microbiology Reviews*, vol. 11, no. 3, pp. 480-96.

Gubler, D.J. 2011, 'Dengue, urbanization and globalization: the unholy trinity of the 21st century', *Tropical Medicine and Health*, vol. 39, pp. S3-S11.

Hafeez, S., Amin, M. & Munir, B.A. 2017, 'Spatial mapping of temporal risk to improve prevention measures: a case study of dengue epidemic in Lahore', *Spatial and Spatio-temporal Epidemiology*, vol. 21, pp. 77-85.

Hagenlocher, M. & Castro, M.C. 2015, 'Mapping malaria risk and vulnerability in the United Republic of Tanzania: a spatial explicit model', *Population Health Metrics*, vol. 13, no. 1, pp. 1-14.

Hales, S., De Wet, N., Maindonald, J. & Woodward, A. 2002, 'Potential effect of population and climate changes on global distribution of dengue fever: an empirical model', *The Lancet*, vol. 360, no. 9336, pp. 830-4.

Halloran, M.E., Auranen, K., Baird, S., Basta, N.E., Bellan, S.E., Brookmeyer, R., Cooper, B.S., DeGruttola, V., Hughes, J.P. & Lessler, J. 2017, 'Simulations for designing and interpreting intervention trials in infectious diseases', *BMC Medicine*, vol. 15, no. 1, pp. 1-8.

Hashem, A.M., Abujamel, T., Alhabbab, R., Almazroui, M. & Azhar, E.I. 2018, 'Dengue infection in patients with febrile illness and its relationship to climate factors: a case study in the city of Jeddah, Saudi Arabia, for the period 2010–2014', *Acta Tropica*, vol. 181, pp. 105-11.

Henley, S.S., Golden, R.M. & Kashner, T.M. 2020, 'Statistical modeling methods: challenges and strategies', *Biostatistics & Epidemiology*, vol. 4, no. 1, pp. 105-39.

Higa, Y. 2011, 'Dengue vectors and their spatial distribution', *Tropical Medicine and Health*, vol. 39, pp. S17-S27.

Hsueh, Y.-H., Lee, J. & Beltz, L. 2012, 'Spatio-temporal patterns of dengue fever cases in Kaoshiung city, Taiwan, 2003–2008', *Applied Geography*, vol. 34, pp. 587-94.

Huang, C.-C., Tam, T.Y.T., Chern, Y.-R., Lung, S.-C.C., Chen, N.-T. & Wu, C.-D. 2018, 'Spatial clustering of dengue fever incidence and its association with surrounding greenness', *International Journal of Environmental Research and Public Health*, vol. 15, no. 9, p. 1869.

Huang, Z., Das, A., Qiu, Y. & Tatem, A.J. 2012, 'Web-based GIS: the vector-borne disease airline importation risk (VBD-AIR) tool', *International Journal of Health Geographics*, vol. 11, no. 1, pp. 1-14.

Humphrey, J.M., Cleton, N.B., Reusken, C.B., Glesby, M.J., Koopmans, M.P. & Abu-Raddad, L.J. 2016, 'Dengue in the Middle East and North Africa: a systematic review', *PLOS Neglected Tropical Diseases*, vol. 10, no. 12, p. e0005194.

Ij, H. 2018, 'Statistics versus machine learning', *Nat Methods*, vol. 15, no. 4, p. 233.

Imrey, P.B. 2000, 'Poisson regression, logistic regression, and loglinear models for random counts', *Handbook of Applied Multivariate Statistics and Mathematical Modeling*, Elsevier, pp. 391-437.

Jácome, G., Vilela, P. & Yoo, C. 2019, 'Social-ecological modelling of the spatial distribution of dengue fever and its temporal dynamics in Guayaquil, Ecuador for climate change adaption', *Ecological Informatics*, vol. 49, pp. 1-12.

Jain, R., Sontisirikit, S., Iamsirithaworn, S. & Prendinger, H. 2019, 'Prediction of dengue outbreaks based on disease surveillance, meteorological and socio-economic data', *BMC Infectious Diseases*, vol. 19, no. 1, pp. 1-16.

Jainul Fathima, A., Revathy, R., Balamurali, S. & Murugaboopathi, G. 2019, 'Prediction of dengue-human protein interaction using artificial neural network for Anti-viral drug discovery', paper presented to the *Proceedings of International Conference on Sustainable Computing in Science, Technology and Management (SUSCOM), Amity University Rajasthan, Jaipur-India*.

Jakobsen, J.C., Gluud, C., Wetterslev, J. & Winkel, P. 2017, 'When and how should multiple imputation be used for handling missing data in randomised clinical trials–a practical guide with flowcharts', *BMC Medical Research Methodology*, vol. 17, no. 1, pp. 1-10.

Jeefoo, P. 2012, 'Spatial temporal dynamics and risk zonation of dengue fever, dengue hemorrhagic fever, and dengue shock syndrome in Thailand', *International Journal of Modern Education & Computer Science*, vol. 4, no. 9.

Jeefoo, P., Tripathi, N.K. & Souris, M. 2011, 'Spatio-temporal diffusion pattern and hotspot detection of dengue in Chachoengsao province, Thailand', *International Journal of Environmental Research and Public Health*, vol. 8, no. 1, pp. 51-74.

Jerez, J.M., Molina, I., García-Laencina, P.J., Alba, E., Ribelles, N., Martín, M. & Franco, L. 2010, 'Missing data imputation using statistical and machine learning methods in a real breast cancer problem', *Artificial Intelligence in Medicine*, vol. 50, no. 2, pp. 105-15.

Jogun, T. 2016, 'The simulation model of land cover change in the Požega-Slavonia County', *Uni-versity of Zagreb. Faculty of Science. Department of Geography*.

Jogun, T., Lukić, A. & Gašparović, M. 2019, 'Simulation model of land cover changes in a post-socialist peripheral rural area: Požega-Slavonia county, Croatia', *Croatian Geographical Bulletin*, vol. 81, no. 1.

Kafy, A.-A., Naim, M.N.H., Subramanyam, G., Ahmed, N.U., Al Rakib, A., Kona, M.A. & Sattar, G.S. 2021, 'Cellular automata approach in dynamic modelling of land cover changes using RapidEye images in Dhaka, Bangladesh', *Environmental Challenges*, vol. 4, p. 100084.

Kamaraj, M. & Rangarajan, S. 2022, 'Predicting the future land use and land cover changes for Bhavani basin, Tamil Nadu, India, using QGIS MOLUSCE plugin', *Environmental Science and Pollution Research*, pp. 1-12.

Kamkhad, N., Jampachaisri, K., Natwichai, J., Siriyasatien, P. & Kesorn, K. 2016, 'Semantic-based data imputation for dengue fever information', paper presented to the *Proceedings of International Workshop on Smart Info-Media Systems in Asia, Ayutthaya, Thailand*.

Kang, H. 2013, 'The prevention and handling of the missing data', *Korean Journal of Anesthesiology*, vol. 64, no. 5, p. 402.

Khalil, H. & Wainer, G. 2020, 'Cell-DEVS for social phenomena modeling', *IEEE Transactions on Computational Social Systems*, vol. 7, no. 3, pp. 725-40.

Khan, M.A. 2021, 'Dengue infection modeling and its optimal control analysis in East Java, Indonesia', *Heliyon*, vol. 7, no. 1, p. e06023.

Khan, N.A., Azhar, E.I., El-Fiky, S., Madani, H.H., Abuljadial, M.A., Ashshi, A.M., Turkistani, A.M. & Hamouh, E.A. 2008, 'Clinical profile and outcome of hospitalized patients during first outbreak of dengue in Makkah, Saudi Arabia', *Acta Tropica*, vol. 105, no. 1, pp. 39-44.

Kholedi, A., Balubaid, O., Milaat, W., Kabbash, I. & Ibrahim, A. 2012, 'Factors associated with the spread of dengue fever in Jeddah governorate, Saudi Arabia', *Eastern Mediterranean Health Journal (EMHJ)*, pp. 15-23.

Khormi, H.M. & Kumar, L. 2011, 'Modeling dengue fever risk based on socioeconomic parameters, nationality and age groups: GIS and remote sensing based case study', *Science of the Total Environment*, vol. 409, no. 22, pp. 4713-9.

Khormi, H.M. & Kumar, L. 2012, 'The importance of appropriate temporal and spatial scales for dengue fever control and management', *Science of the Total Environment*, vol. 430, pp. 144-9.

Khormi, H.M., Kumar, L. & Elzahrany, R.A. 2011, 'Modeling spatio-temporal risk changes in the incidence of dengue fever in Saudi Arabia: a geographical information system case study', *Geospatial Health*, vol. 6, no. 1, pp. 77-84.

Kositsakulchai, E., Phankamolsil, Y. & Yodjaroen, S. 2021, 'Future runoff projections based on land change using integrated Markov-cellular automata model and soil water assessment tool in Lam Pachi Basin, Thailand', *Agriculture and Natural Resources*, vol. 55, no. 5, pp. 810–9-–9.

Koyadun, S., Butraporn, P. & Kittayapong, P. 2012, 'Ecologic and sociodemographic risk determinants for dengue transmission in urban areas in Thailand', *Interdisciplinary Perspectives on Infectious Diseases*, vol. 2012.

Kugler, T.A., Van Riper, D.C., Manson, S.M., Haynes II, D.A., Donato, J. & Stinebaugh, K. 2015, 'Terra Populus: Workflows for integrating and harmonizing geospatial population and environmental data', *Journal of Map & Geography Libraries*, vol. 11, no. 2, pp. 180-206.

Lakshita, N. & Rahayu, S. 2021, 'Urban dynamics and carbon stock estimation in Salatiga city, Indonesia', paper presented to the *IOP Conference Series: Earth and Environmental Science*.

Lary, D.J., Alavi, A.H., Gandomi, A.H. & Walker, A.L. 2016, 'Machine learning in geosciences and remote sensing', *Geoscience Frontiers*, vol. 7, no. 1, pp. 3-10.

Latif, B.A., Mercier, G. & Matsopouls, G. 2010, *Self-organizing maps for processing of data with missing values and outliers: application to remote sensing images*, IntechOpen.

Laureano-Rosario, A.E., Duncan, A.P., Mendez-Lazaro, P.A., Garcia-Rejon, J.E., Gomez-Carro, S., Farfan-Ale, J., Savic, D.A. & Muller-Karger, F.E. 2018, 'Application of artificial neural networks for dengue fever outbreak predictions in

the northwest coast of Yucatan, Mexico and San Juan, Puerto Rico', *Tropical Medicine and Infectious Disease*, vol. 3, no. 1, p. 5.

Lazar, A., Jin, L., Spurlock, C.A., Wu, K. & Sim, A. 2017, 'Data quality challenges with missing values and mixed types in joint sequence analysis', paper presented to the *2017 IEEE International Conference on Big Data (Big Data)*.

Lega, J., Brown, H.E. & Barrera, R. 2017, 'Aedes aegypti (Diptera: Culicidae) abundance model improved with relative humidity and precipitation-driven egg hatching', *Journal of Medical Entomology*, vol. 54, no. 5, pp. 1375-84.

Lemos, C.M.G., de Castro Medeiros, L.C., Ribeiro, K. & Avancini, R. 2017, 'Agent-based model implemented using the TerraME framework to simulate the dynamic transmission of dengue fever', *Revista Geografias*, pp. 85-98.

Li, Y., Dou, Q., Lu, Y., Xiang, H., Yu, X. & Liu, S. 2020, 'Effects of ambient temperature and precipitation on the risk of dengue fever: a systematic review and updated meta-analysis', *Environmental Research*, vol. 191, p. 110043.

Lin, C.-H. & Wen, T.-H. 2011, 'Using geographically weighted regression (GWR) to explore spatial varying relationships of immature mosquitoes and human densities with the incidence of dengue', *International Journal of Environmental Research and Public Health*, vol. 8, no. 7, pp. 2798-815.

Liu, D., Guo, S., Zou, M., Chen, C., Deng, F., Xie, Z., Hu, S. & Wu, L. 2019, 'A dengue fever predicting model based on Baidu search index data and climate data in South China', *PLOS ONE*, vol. 14, no. 12, p. e0226841.

Lloyd, C. 2010, *Spatial data analysis: an introduction for GIS users*, Oxford University Press.

Lukasczyk, J., Maciejewski, R., Garth, C. & Hagen, H. 2015, 'Understanding hotspots: a topological visual analytics approach', paper presented to the *Proceedings of the 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems*.

M Ashshi, A., Alghamdi, S., El-Shemi, A.G., Almdani, S., Refaat, B., Mohamed, A.M., Ghazi, H.O., Azhar, E.I. & Al-Allaf, F.A. 2017, 'Seroprevalence of asymptomatic dengue virus infection and its antibodies among healthy/eligible saudi blood donors: findings from holy Makkah city', *Virology: Research and Treatment*, vol. 8, p. 1178122X17691261.

Machault, V., Yébakima, A., Etienne, M., Vignolles, C., Palany, P., Tourre, Y.M., Guérécheau, M. & Lacaux, J.-P. 2014, 'Mapping entomological dengue risk levels

in Martinique using high-resolution remote-sensing environmental data', *ISPRS International Journal of Geo-Information*, vol. 3, no. 4, pp. 1352-71.

Mala, S. & Jat, M.K. 2019a, 'Geographic information system based spatio-temporal dengue fever cluster analysis and mapping', *The Egyptian Journal of Remote Sensing and Space Science*, vol. 22, no. 3, pp. 297-304.

Mala, S. & Jat, M.K. 2019b, 'Implications of meteorological and physiographical parameters on dengue fever occurrences in Delhi', *Science of the Total Environment*, vol. 650, pp. 2267-83.

Manica, M., Filipponi, F., D'Alessandro, A., Screti, A., Neteler, M., Rosa, R., Solimini, A., Della Torre, A. & Caputo, B. 2016, 'Spatial and temporal hot spots of Aedes albopictus abundance inside and outside a south European metropolitan area', *PLOS Neglected Tropical Diseases*, vol. 10, no. 6, p. e0004758.

Marti, R., Li, Z., Catry, T., Roux, E., Mangeas, M., Handschumacher, P., Gaudart, J., Tran, A., Demagistri, L. & Faure, J.-F. 2020, 'A mapping review on urban landscape factors of dengue retrieved from earth observation data, GIS techniques, and survey questionnaires', *Remote Sensing*, vol. 12, no. 6, p. 932.

Mathur, N., Asirvadam, V.S. & Dass, S.C. 2018, 'Spatial-temporal visualization of dengue incidences using Gaussian Kernel', paper presented to the *2018 International Conference on Intelligent and Advanced System (ICIAS)*.

McGough, S.F., Clemente, L., Kutz, J.N. & Santillana, M. 2021, 'A dynamic, ensemble learning approach to forecast dengue fever epidemic years in Brazil using weather and population susceptibility cycles', *Journal of the Royal Society Interface*, vol. 18, no. 179, p. 20201006.

Medeiros, L.C.d.C., Castilho, C.A.R., Braga, C., de Souza, W.V., Regis, L. & Monteiro, A.M.V. 2011, 'Modeling the dynamic transmission of dengue fever: investigating disease persistence', *PLOS Neglected Tropical Diseases*, vol. 5, no. 1, p. e942.

Mekha, P., Osathanunkul, K. & Teeyasuksaet, N. 2016, 'Gene classification of dengue virus type based on codon usage', paper presented to the *2016 International Computer Science and Engineering Conference (ICSEC)*.

Memarian, H., Balasundram, S.K., Talib, J.B., Sung, C.T.B., Sood, A.M. & Abbaspour, K. 2012, 'Validation of CA-Markov for simulation of land use and cover change in the Langat Basin, Malaysia', *Journal of Geographic Information System*

Memish, Z.A. 2002, 'Infection control in Saudi Arabia: meeting the challenge', *American Journal of Infection Control*, vol. 30, no. 1, pp. 57-65.

Memish, Z.A., Venkatesh, S. & Ahmed, Q.A. 2003, 'Travel epidemiology: the Saudi perspective', *International Journal of Antimicrobial Agents*, vol. 21, no. 2, pp. 96-101.

Meyer, R. 2014, 'Deviance information criterion (DIC)', *Wiley StatsRef: Statistics Reference Online*, pp. 1-6.

Moher, D., Liberati, A., Tetzlaff, J. & Altman, D.G. 2010, 'Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement', *Int J Surg*, vol. 8, no. 5, pp. 336-41.

Moher, D., Liberati, A., Tetzlaff, J., Altman, D.G. & Group*, P. 2009, 'Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement', *Annals of Internal Medicine*, vol. 151, no. 4, pp. 264-9.

Moncayo, A.C., Fernandez, Z., Ortiz, D., Diallo, M., Sall, A., Hartman, S., Davis, C.T., Coffey, L., Mathiot, C.C. & Tesh, R.B. 2004, 'Dengue emergence and adaptation to peridomestic mosquitoes', *Emerging Infectious Diseases*, vol. 10, no. 10, p. 1790.

Mondini, A. & Chiaravalloti-Neto, F. 2008, 'Spatial correlation of incidence of dengue with socioeconomic, demographic and environmental variables in a Brazilian city', *Science of the Total Environment*, vol. 393, no. 2-3, pp. 241-8.

Mordecai, E.A., Cohen, J.M., Evans, M.V., Gudapati, P., Johnson, L.R., Lippi, C.A., Miazgowicz, K., Murdock, C.C., Rohr, J.R. & Ryan, S.J. 2017, 'Detecting the impact of temperature on transmission of zika, dengue, and chikungunya using mechanistic models', *PLOS Neglected Tropical Diseases*, vol. 11, no. 4, p. e0005568.

Mudele, O., Bayer, F.M., Zanandrez, L.F., Eiras, A.E. & Gamba, P. 2020, 'Modeling the temporal population distribution of Ae.aegypti mosquito using big earth observation data', *IEEE Access*, vol. 8, pp. 14182-94.

Murad, A.A. 2007, 'Creating a GIS application for health services at Jeddah city', *Computers in Biology and Medicine*, vol. 37, no. 6, pp. 879-89.

Mutheneni, S.R., Mopuri, R., Naish, S., Gunti, D. & Upadhyayula, S.M. 2018, 'Spatial distribution and cluster analysis of dengue using self organizing maps in Andhra Pradesh, India, 2011–2013', *Parasite Epidemiology and Control*, vol. 3, no. 1, pp. 52-61.

Naish, S., Dale, P., Mackenzie, J.S., McBride, J., Mengersen, K. & Tong, S. 2014a, 'Climate change and dengue: a critical and systematic review of quantitative modelling approaches', *BMC Infectious Diseases*, vol. 14, no. 1, pp. 1-14.

Naish, S., Dale, P., Mackenzie, J.S., McBride, J., Mengersen, K. & Tong, S. 2014b, 'Spatial and temporal patterns of locally-acquired dengue transmission in northern Queensland, Australia, 1993–2012', *PLOS ONE*, vol. 9, no. 4, p. e92524.

Naish, S. & Tong, S. 2014, 'Hot spot detection and spatio-temporal dynamics of dengue in Queensland, Australia', paper presented to the *Proceedings of the ISPRS Technical Commission VIII Symposium [International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences-ISPRS Archives*.

Nakhapakorn, K. & Tripathi, N.K. 2005, 'An information value based analysis of physical and climatic factors affecting dengue fever and dengue haemorrhagic fever incidence', *International Journal of Health Geographics*, vol. 4, no. 1, pp. 1-13.

Nakvisut, A. & Phienthrakul, T. 2018, 'Two-step prediction technique for dengue outbreak in Thailand', paper presented to the *2018 International Electrical Engineering Congress (iEECON)*.

Nazeer, M. & Bilal, M. 2018, 'Evaluation of ordinary least square (OLS) and geographically weighted regression (GWR) for water quality monitoring: a case study for the estimation of salinity', *Journal of Ocean University of China*, vol. 17, no. 2, pp. 305-10.

Ngueilbaye, A., Wang, H., Mahamat, D.A. & Junaidu, S.B. 2021, 'Modulo 9 model-based learning for missing data imputation', *Applied Soft Computing*, vol. 103, p. 107167.

Nguyen, L.T., Le, H.X., Nguyen, D.T., Ho, H.Q. & Chuang, T.-W. 2020, 'Impact of climate variability and abundance of mosquitoes on dengue transmission in central Vietnam', *International Journal of Environmental Research and Public Health*, vol. 17, no. 7, p. 2453.

Nisha, M., Mohanavalli, S. & Swathika, R. 2013, 'Improving the quality of clustering using cluster ensembles', paper presented to the *2013 IEEE Conference on Information & Communication Technologies*.

Nithyaa, K., Anandhasreb, R., Anusuyab, D., Moogabigaib, B. & Reshmab, K. 2019, 'Dengue diseases prediction using SMO classification', *South Asian Journal of Engineering and Technology*, vol. 8, no. 01, pp. 88-91.

Onan, A. 2015, 'A fuzzy-rough nearest neighbor classifier combined with consistency-based subset evaluation and instance selection for automated diagnosis of breast cancer', *Expert Systems with Applications*, vol. 42, no. 20, pp. 6844-52.

Onan, A. 2019, 'Consensus clustering-based undersampling approach to imbalanced learning', *Scientific Programming*, vol. 2019.

Onan, A. 2020, 'Mining opinions from instructor evaluation reviews: a deep learning approach', *Computer Applications in Engineering Education*, vol. 28, no. 1, pp. 117-38.

Onan, A., Korukoğlu, S. & Bulut, H. 2016, 'Ensemble of keyword extraction methods and classifiers in text classification', *Expert Systems with Applications*, vol. 57, pp. 232-47.

Ong, J., Liu, X., Rajarethinam, J., Kok, S.Y., Liang, S., Tang, C.S., Cook, A.R., Ng, L.C. & Yap, G. 2018, 'Mapping dengue risk in Singapore using random forest', *PLOS Neglected Tropical Diseases*, vol. 12, no. 6, p. e0006587.

Organji, S.R., Abulreesh, H.H. & Osman, G.E. 2017, 'Circulation of dengue virus serotypes in the city of Makkah, Saudi Arabia, as determined by reverse transcription polymerase chain reaction', *Canadian Journal of Infectious Diseases and Medical Microbiology*, vol. 2017.

Ortigoza, G., Brauer, F. & Lorandi, A. 2019, 'Mosquito-borne diseases simulated by cellular automata: a review', *International Journal of Mosquito Research*, vol. 6, no. 6, pp. 31-8.

Ortiz, P.L., Rivero, A., Linares, Y., Pérez, A. & Vázquez, J.R. 2015, 'Spatial models for prediction and early warning of Aedes aegypti proliferation from data on climate change and variability in Cuba', *MEDICC Review*, vol. 17, pp. 20-8.

Papenmeier, A., Englebienne, G. & Seifert, C. 2019, 'How model accuracy and explanation fidelity influence user trust', *arXiv preprint arXiv:1907.12652*.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R. & Dubourg, V. 2011, 'Scikit-learn: machine learning in Python', *The Journal of Machine Learning Research*, vol. 12, pp. 2825-30.

Pereira, F. & Schimit, P. 2018, 'Dengue fever spreading based on probabilistic cellular automata with two lattices', *Physica A: Statistical Mechanics and its Applications*, vol. 499, pp. 75-87.

Pereira, F.H., Schimit, P.H. & Bezerra, F.E. 2021, 'A deep learning based surrogate model for the parameter identification problem in probabilistic cellular automaton epidemic models', *Computer Methods and Programs in Biomedicine*, vol. 205, p. 106078.

Phakhounthong, K., Chaovalit, P., Jittamala, P., Blacksell, S.D., Carter, M.J., Turner, P., Chheng, K., Sona, S., Kumar, V. & Day, N.P. 2018, 'Predicting the severity of dengue fever in children on admission based on clinical features and laboratory indicators: application of classification tree analysis', *BMC Pediatrics*, vol. 18, no. 1, pp. 1-9.

Philemon, M.D., Ismail, Z. & Dare, J. 2019, 'A review of epidemic forecasting using artificial neural networks', *International Journal of Epidemiologic Research*, vol. 6, no. 3, pp. 132-43.

Phuyal, P., Kramer, I.M., Klingelhöfer, D., Kuch, U., Madeburg, A., Groneberg, D.A., Wouters, E., Dhimal, M. & Müller, R. 2020, 'Spatiotemporal distribution of dengue and chikungunya in the Hindu Kush Himalayan region: a systematic review', *International Journal of Environmental Research and Public Health*, vol. 17, no. 18, p. 6656.

Pigott, T.D. 2001, 'A review of methods for missing data', *Educational Research and Evaluation*, vol. 7, no. 4, pp. 353-83.

Powell, J.R. & Tabachnick, W.J. 2013, 'History of domestication and spread of Aedes aegypti-a review', *Mem Inst Oswaldo Cruz*, vol. 108, pp. 11-7.

Prasetyowati, H., Dhewantara, P.W., Hendri, J., Astuti, E.P., Gelaw, Y.A., Harapan, H., Ipa, M., Widyastuti, W., Handayani, D.O.T.L. & Salama, N. 2021, 'Geographical heterogeneity and socio-ecological risk profiles of dengue in Jakarta, Indonesia', *Geospatial Health*, vol. 16, no. 1.

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A.V. & Gulin, A. 2018, 'CatBoost: unbiased boosting with categorical features', *Advances in Neural Information Processing Systems*, vol. 31.

Puggioni, G., Couret, J., Serman, E., Akanda, A.S. & Ginsberg, H.S. 2020, 'Spatiotemporal modeling of dengue fever risk in Puerto Rico', *Spatial and Spatio-temporal Epidemiology*, vol. 35, p. 100375.

Qi, X., Wang, Y., Li, Y., Meng, Y., Chen, Q., Ma, J. & Gao, G.F. 2015, 'The effects of socioeconomic and environmental factors on the incidence of dengue fever in the

Pearl River Delta, China, 2013', *PLOS Neglected Tropical Diseases*, vol. 9, no. 10, p. e0004159.

Racloz, V., Ramsey, R., Tong, S. & Hu, W. 2012, 'Surveillance of dengue fever virus: a review of epidemiological models and early warning systems', *PLOS Neglected Tropical Diseases*, vol. 6, no. 5, p. e1648.

Rahman, F. & Rahman, M.T.U. 2021, 'Use of cellular automata-based artificial neural networks for detection and prediction of land use changes in north eastern Dhaka city', *Research Square*.

Rahman, M., Tabassum, F., Rasheduzzaman, M., Saba, H., Sarkar, L., Ferdous, J., Uddin, S.Z. & Zahedul Islam, A. 2017, 'Temporal dynamics of land use/land cover change and its prediction using CA-ANN model for southwestern coastal Bangladesh', *Environmental Monitoring and Assessment*, vol. 189, no. 11, pp. 1-18.

Rahman, S.A., Rahim, A. & Mallongi, A. 2018, 'Risk analysis of dengue fever occurrence in bone province sulawesi south using temporal spatial geostatistical model', *Indian Journal of Public Health Research & Development*, vol. 9, no. 4, pp. 221-6.

Raju, N.G., Krishna, P.G., Manognya, K., Kiran, G.R., Rohit, P. & Likhith, K. 2019, 'Evolution of predictive model for dengue incidence by using machine learning algorithms', paper presented to the *2019 International Conference on Communication and Electronics Systems (ICCES)*.

Ramesh, J., Aburukba, R. & Sagahyroon, A. 2021, 'A remote healthcare monitoring framework for diabetes prediction using machine learning', *Healthcare Technology Letters*, vol. 8, no. 3, p. 45.

Rao, K.V., Govardhan, A. & Rao, K.C. 2012, 'Spatiotemporal data mining: issues, tasks and applications', *International Journal of Computer Science and Engineering Survey*, vol. 3, no. 1, p. 39.

Reiter, P. 2001, 'Climate change and mosquito-borne disease', *Environmental Health Perspectives*, vol. 109, pp. 141-61.

Ren, H., Wu, W., Li, T. & Yang, Z. 2019, 'Urban villages as transfer stations for dengue fever epidemic: a case study in the Guangzhou, China', *PLOS Neglected Tropical Diseases*, vol. 13, no. 4, p. e0007350.

Ren, H., Zheng, L., Li, Q., Yuan, W. & Lu, L. 2017, 'Exploring determinants of spatial variations in the dengue fever epidemic using geographically weighted regression

model: a case study in the joint Guangzhou-Foshan area, China, 2014', *International Journal of Environmental Research and Public Health*, vol. 14, no. 12, p. 1518.

Restrepo, A.C., Baker, P. & Clements, A.C. 2014, 'National spatial and temporal patterns of notified dengue cases, Colombia 2007–2010', *Tropical Medicine & International Health*, vol. 19, no. 7, pp. 863-71.

Ribeiro, M.T., Singh, S. & Guestrin, C. 2016, 'Why should I trust you? Explaining the predictions of any classifier', paper presented to the *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Richman, M.B., Trafalis, T.B. & Adrianto, I. 2009, 'Missing data imputation through machine learning algorithms', *Artificial Intelligence Methods in the Environmental Sciences*, Springer, pp. 153-69.

Rose, A., McKee, J., Urban, M. & Bright, E. 2018, 'LandScan Global 2017', electronic data set, <landscan.ornl.gov>.

Rose, A., McKee, J., Urban, M., Bright, E. & Sims, K. 2019, 'LandScan Global 2018', electronic data set, <landscan.ornl.gov>.

Roslan, N.S., Abd Latif, Z. & Dom, N.C. 2016, 'Dengue cases distribution based on land surface temperature and elevation', paper presented to the *2016 7th IEEE Control and System Graduate Research Colloquium (ICSGRC)*.

Rotela, C., Fouque, F., Lamfri, M., Sabatier, P., Introini, V., Zaidenberg, M. & Scavuzzo, C. 2007, 'Space–time analysis of the dengue spreading dynamics in the 2004 Tartagal outbreak, Northern Argentina', *Acta Tropica*, vol. 103, no. 1, pp. 1-13.

Rubin, D.B. 1976, 'Inference and missing data', *Biometrika*, vol. 63, no. 3, pp. 581-92.

Runge-Ranzinger, S., McCall, P.J., Kroeger, A. & Horstick, O. 2014, 'Dengue disease surveillance: an updated systematic literature review', *Tropical Medicine & International Health*, vol. 19, no. 9, pp. 1116-60.

Saad, A. 2017, 'Health issues during Hajj', *The Egyptian Journal of Internal Medicine*, vol. 29, no. 2, pp. 37-9.

Saha, S., Moorthi, S., Pan, H.-L., Wu, X., Wang, J., Nadiga, S., Tripp, P., Kistler, R., Woollen, J. & Behringer, D. 2010, 'The NCEP climate forecast system reanalysis', *Bulletin of the American Meteorological Society*, Series The NCEP climate forecast system reanalysis vol. 91, no 8, pp. 1015-58.

Saha, S., Moorthi, S., Wu, X., Wang, J., Nadiga, S., Tripp, P., Behringer, D., Hou, Y.-T., Chuang, H.-y. & Iredell, M. 2011, 'NCEP climate forecast system version 2

(CFSv2) 6-hourly products', *Research Data Archive at the National Center for Atmospheric Research, Computational and Information Systems Laboratory*, vol. 10, p. D61C1TXF.

Saha, T.K., Pal, S. & Sarkar, R. 2021, 'Prediction of wetland area and depth using linear regression model and artificial neural network based cellular automata', *Ecological Informatics*, vol. 62, p. 101272.

Sajana, T., Navya, M., Gayathri, Y. & Reshma, N. 2018, 'Classification of dengue using machine learning techniques', *Int J Eng Technol*, vol. 7, no. 2.32, pp. 212-8.

Sallam, M.F., Fizer, C., Pilant, A.N. & Whung, P.-Y. 2017, 'Systematic review: land cover, meteorological, and socioeconomic determinants of Aedes mosquito habitat for risk mapping', *International Journal of Environmental Research and Public Health*, vol. 14, no. 10, p. 1230.

Santos, L., Costa, M., Pinho, S.T.R.d., Andrade, R.F.S., Barreto, F.R., Teixeira, M. & Barreto, M.L. 2009, 'Periodic forcing in a three-level cellular automata model for a vector-transmitted disease', *Physical Review E*, vol. 80, no. 1, p. 016102.

Santos, L.B.L., Maretto, R.V., de Castro Medeiros, L.C., da Fonseca Feitosa, F. & Monteiro, A.M.V. 2011, 'A susceptible-infected model for exploring the effects of neighborhood structures on epidemic processes-a segregation analysis', paper presented to the *GeoInfo*.

Sarfraz, M.S., Tripathi, N.K. & Kitamoto, A. 2014, 'Near real-time characterisation of urban environments: a holistic approach for monitoring dengue fever risk areas', *International Journal of Digital Earth*, vol. 7, no. 11, pp. 916-34.

Sarma, D., Hossain, S., Mittra, T., Bhuiya, M.A.M., Saha, I. & Chakma, R. 2020, 'Dengue prediction using machine learning algorithms', paper presented to the *2020 IEEE 8th R10 Humanitarian Technology Conference (R10-HTC)*.

Saxena, A., Prasad, M., Gupta, A., Bharill, N., Patel, O.P., Tiwari, A., Er, M.J., Ding, W. & Lin, C.-T. 2017, 'A review of clustering techniques and developments', *Neurocomputing*, vol. 267, pp. 664-81.

Scavuzzo, J.M., Trucco, F., Espinosa, M., Tauro, C.B., Abril, M., Scavuzzo, C.M. & Frery, A.C. 2018, 'Modeling dengue vector population using remotely sensed data and machine learning', *Acta Tropica*, vol. 185, pp. 167-75.

Scheffer, J. 2002, 'Dealing with missing data', *Mathematical Research Letters*.

Schmidt, W.-P., Suzuki, M., Dinh Thiem, V., White, R.G., Tsuzuki, A., Yoshida, L.-M., Yanai, H., Haque, U., Huu Tho, L. & Anh, D.D. 2011, 'Population density, water

supply, and the risk of dengue fever in Vietnam: cohort study and spatial analysis', *PLOS Medicine*, vol. 8, no. 8, p. e1001082.

Schwalbert, R.A., Amado, T., Corassa, G., Pott, L.P., Prasad, P.V. & Ciampitti, I.A. 2020, 'Satellite-based soybean yield forecast: integrating machine learning and weather data for improving crop yield prediction in southern Brazil', *Agricultural and Forest Meteorology*, vol. 284, p. 107886.

Sessa, J. & Syed, D. 2016, 'Techniques to deal with missing data', paper presented to the *2016 5th International Conference on Electronic Devices, Systems and Applications (ICEDSA)*.

Shafi, S., Memish, Z.A., Gatrad, A.R. & Sheikh, A. 2005, 'Hajj 2006: communicable disease and other health risks and current official guidance for pilgrims', *Euro Surveill*, vol. 10, no. 50, p. 2857.

Shahina, W., Nassara, A., Kalkattawia, M. & Bokharia, H. 2009, 'Dengue fever in a tertiary hospital in Makkah, Saudi Arabia', *WHO Regional Office for South-East Asia*.

Shi, Z. & Pun-Cheng, L.S. 2019, 'Spatiotemporal data clustering: a survey of methods', *ISPRS International Journal of Geo-information*, vol. 8, no. 3, p. 112.

Shukla, N., Hagenbuchner, M., Win, K.T. & Yang, J. 2018, 'Breast cancer data analysis for survivability studies and prediction', *Computer Methods and Programs in Biomedicine*, vol. 155, pp. 199-208.

Silva, L.O. & Zárate, L.E. 2014, 'A brief review of the main approaches for treatment of missing data', *Intelligent Data Analysis*, vol. 18, no. 6, pp. 1177-98.

Sippy, R., Herrera, D., Gaus, D., Gangnon, R.E., Patz, J.A. & Osorio, J.E. 2019, 'Seasonal patterns of dengue fever in rural Ecuador: 2009-2016', *PLOS Neglected Tropical Diseases*, vol. 13, no. 5, p. e0007360.

Snoek, J., Larochelle, H. & Adams, R.P. 2012, 'Practical bayesian optimization of machine learning algorithms', *Advances in Neural Information Processing Systems*, vol. 25.

Soley-Bori, M. 2013, 'Dealing with missing data: Key assumptions and methods for applied analysis', *Boston University*, vol. 4, no. 1, p. 19.

Sommer, D., Grimm, T. & Golz, M. 2003, 'Processing missing values with self-organized maps', paper presented to the *European Symposium on Intelligent Technologies, Hybrid Systems and their Implementation on Smart Adaptive Systems*.

Sorjamaa, A., Merlin, P., Maillet, B. & Lendasse, A. 2007, 'SOM+ EOF for finding missing values', paper presented to the *ESANN*.

Sriklin, T., Kajornkasirat, S. & Puttinaovarat, S. 2021, 'Dengue transmission mapping with weather-based predictive model in three southernmost Provinces of Thailand', *Sustainability*, vol. 13, no. 12, p. 6754.

Srivastava, S., Soman, S., Rai, A. & Cheema, A.S. 2020, 'An online learning approach for dengue fever classification', paper presented to the *2020 IEEE 33rd International Symposium on Computer-Based Medical Systems (CBMS)*.

Stanforth, A., Moreno-Madriñán, M.J. & Ashby, J. 2016, 'Exploratory analysis of dengue fever niche variables within the Río Magdalena watershed', *Remote Sensing*, vol. 8, no. 9, p. 770.

Sun, G.-Q., Liu, Q.-X., Jin, Z., Chakraborty, A. & Li, B.-L. 2010, 'Influence of infection rate and migration on extinction of disease in spatial epidemics', *Journal of Theoretical Biology*, vol. 264, no. 1, pp. 95-103.

Taghikhani, R. 2020, 'Mathematics of Dengue Transmission Dynamics and Assessment of Wolbachia-Based Interventions', Arizona State University.

Tedeschi, L.O. 2006, 'Assessment of the adequacy of mathematical models', *Agricultural Systems*, vol. 89, no. 2-3, pp. 225-47.

Tedrow, C.A. 2010, 'Using remote sensing, ecological niche modeling, and Geographic Information Systems for Rift Valley fever risk assessment in the United States', George Mason University.

Teurlai, M., Menkès, C.E., Cavarero, V., Degallier, N., Descloux, E., Grangeon, J.-P., Guillaumot, L., Libourel, T., Lucio, P.S. & Mathieu-Daudé, F. 2015, 'Socio-economic and climate factors associated with dengue fever spatial heterogeneity: a worked example in New Caledonia', *PLOS Neglected Tropical Diseases*, vol. 9, no. 12, p. e0004211.

Theodorakos, K., Broeckhove, J. & Willem, L. 2017, 'Examination of influencing factors and high-risk regions of dengue in Nicaragua, using spatiotemporal compartmental simulations', *Tropical Medicine & International Health*, vol. 22, pp. 156-7.

Tian, H., Huang, S., Zhou, S., Bi, P., Yang, Z., Li, X., Chen, L., Cazelles, B., Yang, J. & Luo, L. 2016, 'Surface water areas significantly impacted 2014 dengue outbreaks in Guangzhou, China', *Environmental Research*, vol. 150, pp. 299-305.

Tran, A., Mangeas, M., Demarchi, M., Roux, E., Degenne, P., Haramboure, M., Le Goff, G., Damiens, D., Gouagna, L.-C. & Herbreteau, V. 2020, 'Complementarity of empirical and process-based approaches to modelling mosquito population dynamics with Aedes albopictus as an example—Application to the development of an operational mapping tool of vector populations', *PLOS ONE*, vol. 15, no. 1, p. e0227407.

Tu, J. & Xia, Z.-G. 2008, 'Examining spatially varying relationships between land use and water quality using geographically weighted regression I: model design and evaluation', *Science of the Total Environment*, vol. 407, no. 1, pp. 358-78.

Ullah, S., Tahir, A.A., Akbar, T.A., Hassan, Q.K., Dewan, A., Khan, A.J. & Khan, M. 2019, 'Remote sensing-based quantification of the relationships between land use land cover changes and surface temperature over the lower Himalayan region', *Sustainability*, vol. 11, no. 19, p. 5492.

Valdez, L.D., Sibona, G.J. & Condat, C. 2018, 'Impact of rainfall on Aedes aegypti populations', *Ecological Modelling*, vol. 385, pp. 96-105.

Valles, J., Perez, C. & Blanco, A. 2019, 'Geospatial and clustering analysis of dengue cases using self-organizing maps: case of Quezon city, 2010–2015', *International Archives of the Photogrammetry, Remote Sensing & Spatial Information Sciences*.

Vesanto, J. & Alhoniemi, E. 2000, 'Clustering of the self-organizing map', *IEEE Transactions on Neural Networks*, vol. 11, no. 3, pp. 586-600.

Viennet, E., Ritchie, S.A., Faddy, H.M., Williams, C.R. & Harley, D. 2014, 'Epidemiology of dengue in a high-income country: a case study in Queensland, Australia', *Parasites & Vectors*, vol. 7, no. 1, pp. 1-16.

Vincenti-Gonzalez, M.F., Grillet, M.-E., Velasco-Salas, Z.I., Lizarazo, E.F., Amarista, M.A., Sierra, G.M., Comach, G. & Tami, A. 2017, 'Spatial analysis of dengue seroprevalence and modeling of transmission risk factors in a dengue hyperendemic city of Venezuela', *PLOS Neglected Tropical Diseases*, vol. 11, no. 1, p. e0005317.

Wang, L. & Wang, G. 2015, 'Data mining applications in big data', *Computer Engineering and Applications Journal*, vol. 4, no. 3, pp. 143-52.

Wen, T.-H., Lin, M.-H., Teng, H.-J. & Chang, N.-T. 2015, 'Incorporating the human-Aedes mosquito interactions into measuring the spatial risk of urban dengue fever', *Applied Geography*, vol. 62, pp. 256-66.

Wen, T.-H., Lin, N.H., Lin, C.-H., King, C.-C. & Su, M.-D. 2006, 'Spatial mapping of temporal risk characteristics to improve environmental health risk identification: a case study of a dengue epidemic in Taiwan', *Science of the Total Environment*, vol. 367, no. 2-3, pp. 631-40.

Wen, T.-H. & Tsai, C.-T. 2016, 'Evaluating the role of disease importation in the spatiotemporal transmission of indigenous dengue outbreak', *Applied Geography*, vol. 76, pp. 137-46.

Wenbai, C., Chang, L., Weizhao, C., Huixiang, L., Qili, C. & Peiliang, W. 2021, 'A prediction method for the RUL of equipment for missing data', *Complexity*, vol. 2021.

Whitehead, S.S., Blaney, J.E., Durbin, A.P. & Murphy, B.R. 2007, 'Prospects for a dengue virus vaccine', *Nature Reviews Microbiology*, vol. 5, no. 7, pp. 518-28.

Whiteman, A., Desjardins, M.R., Eskildsen, G.A. & Loaiza, J.R. 2019, 'Detecting space-time clusters of dengue fever in Panama after adjusting for vector surveillance data', *PLOS Neglected Tropical Diseases*, vol. 13, no. 9, p. e0007266.

Wiese, D., Escalante, A.A., Murphy, H., Henry, K.A. & Gutierrez-Velez, V.H. 2019, 'Integrating environmental and neighborhood factors in MaxEnt modeling to predict species distributions: a case study of Aedes albopictus in southeastern Pennsylvania', *PLOS ONE*, vol. 14, no. 10, p. e0223821.

Wilder-Smith, A. & Gubler, D.J. 2008, 'Geographic expansion of dengue: the impact of international travel', *Medical Clinics of North America*, vol. 92, no. 6, pp. 1377-90.

Wisniewski, S.R., Leon, A.C., Otto, M.W. & Trivedi, M.H. 2006, 'Prevention of missing data in clinical research studies', *Biological Psychiatry*, vol. 59, no. 11, pp. 997-1000.

Wondrade, N., Dick, Ø.B. & Tveite, H. 2014, 'GIS based mapping of land cover changes utilizing multi-temporal remotely sensed image data in Lake Hawassa Watershed, Ethiopia', *Environmental Monitoring and Assessment*, vol. 186, no. 3, pp. 1765-80.

World Health Organization 2017, *Vector-borne diseases*, viewed 22/03/2020, <https://www.who.int/news-room/fact-sheets/detail/vector-borne-diseases>.

World Health Organization 2019, *Dengue and severe dengue*, viewed 19/09/2019, <https://www.who.int/news-room/fact-sheets/detail/dengue-and-severe-dengue>.

Wu, J. & Cowling, B. 2018, 'Real-time forecasting of infectious disease epidemics', *Hong Kong Medical Journal*.

Wu, P.-C., Lay, J.-G., Guo, H.-R., Lin, C.-Y., Lung, S.-C. & Su, H.-J. 2009, 'Higher temperature and urbanization affect the spatial patterns of dengue fever transmission in subtropical Taiwan', *Science of the Total Environment*, vol. 407, no. 7, pp. 2224-33.

Wu, X., Zhu, X., Wu, G.-Q. & Ding, W. 2013, 'Data mining with big data', *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 1, pp. 97-107.

Xiang, J., Hansen, A., Liu, Q., Liu, X., Tong, M.X., Sun, Y., Cameron, S., Hanson-Easey, S., Han, G.-S. & Williams, C. 2017, 'Association between dengue fever incidence and meteorological factors in Guangzhou, China, 2005–2014', *Environmental Research*, vol. 153, pp. 17-26.

Xu, C. & Jackson, S.A. 2019, 'Machine learning and complex biological data', *Genome Biology*, vol. 20, no. 1, pp. 1-4.

Xu, Z., Bambrick, H., Yakob, L., Devine, G., Lu, J., Frentiu, F.D., Yang, W., Williams, G. & Hu, W. 2019, 'Spatiotemporal patterns and climatic drivers of severe dengue in Thailand', *Science of the Total Environment*, vol. 656, pp. 889-901.

Yañez-Arenas, C., Rioja-Nieto, R., Martín, G.A., Dzul-Manzanilla, F., Chiappa-Carrara, X., Buenfil-Ávila, A., Manrique-Saide, P., Correa-Morales, F., Díaz-Quiñónez, J.A. & Pérez-Rentería, C. 2018, 'Characterizing environmental suitability of Aedes albopictus (Diptera: Culicidae) in Mexico based on regional and global niche models', *Journal of Medical Entomology*, vol. 55, no. 1, pp. 69-77.

Yu, H.-L., Yang, S.-J., Yen, H.-J. & Christakos, G. 2011, 'A spatio-temporal climate-based model of early dengue fever warning in southern Taiwan', *Stochastic Environmental Research and Risk Assessment*, vol. 25, no. 4, pp. 485-94.

Yue, Y., Sun, J., Liu, X., Ren, D., Liu, Q., Xiao, X. & Lu, L. 2018, 'Spatial analysis of dengue fever and exploration of its environmental and socio-economic risk factors using ordinary least squares: a case study in five districts of Guangzhou city, China, 2014', *International Journal of Infectious Diseases*, vol. 75, pp. 39-48.

Yung, C.-F., Lee, K.-S., Thein, T.-L., Tan, L.-K., Gan, V.C., Wong, J.G., Lye, D.C., Ng, L.-C. & Leo, Y.-S. 2015, 'Dengue serotype-specific differences in clinical manifestation, laboratory parameters and risk of severe disease in adults, Singapore', *The American Journal of Tropical Medicine and Hygiene*, vol. 92, no. 5, p. 999.

Zaki, A., Perera, D., Jahan, S.S. & Cardosa, M.J. 2008, 'Phylogeny of dengue viruses circulating in Jeddah, Saudi Arabia: 1994 to 2006', *Tropical Medicine & International Health*, vol. 13, no. 4, pp. 584-92.

Zellweger, R.M., Cano, J., Mangeas, M., Taglioni, F., Mercier, A., Despinoy, M., Menkès, C.E., Dupont-Rouzeyrol, M., Nikolay, B. & Teurlai, M. 2017, 'Socioeconomic and environmental determinants of dengue transmission in an urban setting: an ecological study in Nouméa, New Caledonia', *PLOS Neglected Tropical Diseases*, vol. 11, no. 4, p. e0005471.

Zhang, Z. 2016, 'Missing data imputation: focusing on single imputation', *Annals of Translational Medicine*, vol. 4, no. 1.

Zheng, L., Ren, H.-Y., Shi, R.-H. & Lu, L. 2019, 'Spatiotemporal characteristics and primary influencing factors of typical dengue fever epidemics in China', *Infectious Diseases of Poverty*, vol. 8, no. 1, pp. 1-12.

Zhou, S., Zhou, S., Liu, L., Zhang, M., Kang, M., Xiao, J. & Song, T. 2019, 'Examining the effect of the environment and commuting flow from/to epidemic areas on the spread of dengue fever', *International Journal of Environmental Research and Public Health*, vol. 16, no. 24, p. 5013.

Zhu, G., Liu, J., Tan, Q. & Shi, B. 2016, 'Inferring the spatio-temporal patterns of dengue transmission from surveillance data in Guangzhou, China', *PLOS Neglected Tropical Diseases*, vol. 10, no. 4, p. e0004633.

Zhu, G., Xiao, J., Zhang, B., Liu, T., Lin, H., Li, X., Song, T., Zhang, Y., Ma, W. & Hao, Y. 2018, 'The spatiotemporal transmission of dengue and its driving mechanism: a case study on the 2014 dengue outbreak in Guangdong, China', *Science of the Total Environment*, vol. 622, pp. 252-9.

APPENDICES

Appendix A: Supplemental materials to Chapter 4 – Objective 2

Table A.1. Comparison of models performance using annual clusters data identified by applying SOFM and DBSCAN approaches.

			Model													
			Decision Tree		KNN		Random Forest		AdaBoost		SVC		CatBoost		Naive Bayes	
	Cluster	Cluster Size	Training (%)	Testing (%)	Training (%)	Testing (%)	Training (%)	Testing (%)	Training (%)	Testing (%)	Training (%)	Testing (%)	Training (%)	Testing (%)	Training (%)	Testing (%)
2012	Default Parameters															
	-1	133	100%	65%	100%	63%	75%	50%	81%	75%	59%	80%	97%	70%	58%	30%
	0	117	100%	58%	100%	67%	68%	58%	84%	44%	56%	50%	94%	42%	68%	58%
	1	385	100%	65%	100%	77%	67%	69%	74%	67%	57%	55%	83%	72%	56%	62%
	2	1042	100%	66%	100%	67%	74%	77%	71%	76%	54%	54%	89%	78%	56%	57%
	3	198	100%	63%	100%	72%	71%	72%	78%	72%	62%	65%	88%	65%	61%	70%
	4	38	100%	58%	100%	83%	85%	92%	92%	67%	58%	67%	92%	58%	73%	75%
	5	14	100%	20%	100%	20%	100%	0%	100%	20%	67%	20%	100%	40%	44%	20%
	All Clusters	1927	100%	61%	100%	63%	70%	69%	72%	66%	54%	51%	87%	70%	54%	53%
	Hyperparameters															
	-1	133	67%	78%	100%	65%	98%	65%	100%	68%	96%	70%	85%	73%	56%	28%
	0	117	100%	47%	100%	64%	85%	56%	100%	53%	95%	61%	89%	44%	67%	50%
	1	385	62%	57%	100%	78%	90%	77%	100%	62%	71%	66%	82%	71%	61%	64%
	2	1042	70%	66%	74%	64%	77%	68%	100%	66%	79%	67%	84%	79%	53%	53%
	3	198	73%	73%	100%	68%	76%	70%	100%	62%	83%	73%	85%	68%	62%	70%
	4	38	88%	67%	77%	83%	88%	67%	100%	58%	81%	75%	92%	67%	62%	83%
	5	14	67%	20%	67%	20%	67%	20%	100%	0%	67%	20%	100%	40%	67%	20%
	All Clusters	1927	73%	61%	100%	63%	75%	64%	100%	58%	82%	62%	82%	70%	54%	52%
2013	Default Parameters															
	-1	28	100%	33%	100%	33%	100%	33%	100%	33%	79%	22%	100%	33%	84%	33%
	0	4026	100%	74%	100%	76%	79%	77%	82%	82%	71%	72%	89%	82%	71%	71%
	1	791	100%	75%	100%	80%	79%	70%	83%	83%	75%	79%	91%	82%	75%	77%
	2	1017	98%	70%	98%	71%	66%	65%	74%	73%	57%	55%	86%	78%	57%	57%
	3	316	100%	54%	100%	65%	57%	51%	75%	72%	60%	61%	84%	71%	57%	49%
	4	202	99%	80%	99%	77%	70%	77%	89%	77%	76%	70%	94%	77%	83%	72%
	All Clusters	6380	100%	74%	100%	75%	77%	78%	80%	79%	69%	69%	86%	81%	69%	68%
	Hyperparameters															

	-1	28	100%	33%	100%	33%	95%	44%	100%	22%	100%	33%	100%	33%	84%	33%
	0	4026	79%	77%	100%	78%	78%	77%	100%	75%	84%	80%	87%	82%	71%	72%
	1	791	78%	79%	83%	80%	88%	81%	100%	71%	89%	79%	87%	83%	75%	77%
	2	1017	68%	64%	75%	74%	78%	78%	98%	72%	78%	75%	80%	78%	57%	57%
	3	316	71%	68%	75%	67%	67%	64%	100%	59%	80%	71%	77%	72%	57%	49%
	4	202	89%	79%	84%	77%	79%	70%	99%	69%	84%	74%	91%	77%	76%	70%
	All Clusters	6380	80%	79%	100%	76%	84%	79%	100%	73%	83%	77%	84%	81%	69%	69%
2014	**Default Parameters**															
	-1	96	100%	62%	100%	55%	79%	62%	85%	69%	67%	72%	97%	66%	70%	59%
	0	89	100%	44%	100%	48%	74%	52%	77%	52%	60%	59%	85%	48%	63%	48%
	1	729	99%	66%	99%	74%	59%	52%	79%	78%	79%	78%	80%	78%	21%	23%
	2	1982	100%	56%	100%	62%	59%	57%	65%	63%	63%	63%	81%	64%	53%	54%
	3	252	99%	51%	99%	55%	68%	59%	69%	61%	61%	61%	84%	62%	44%	38%
	4	832	98%	59%	98%	58%	65%	60%	65%	61%	56%	55%	81%	69%	52%	48%
	5	132	100%	58%	100%	60%	65%	58%	68%	68%	63%	55%	93%	65%	64%	58%
	6	40	100%	67%	100%	75%	89%	67%	89%	67%	86%	67%	96%	67%	86%	67%
	7	14	100%	40%	100%	40%	89%	40%	100%	20%	67%	80%	100%	40%	67%	60%
	8	20	100%	100%	100%	83%	93%	100%	100%	100%	71%	83%	100%	100%	64%	83%
	9	14	89%	60%	89%	20%	78%	60%	89%	40%	67%	80%	89%	60%	78%	60%
	All Clusters	4200	99%	59%	99%	61%	58%	58%	65%	66%	64%	65%	80%	68%	59%	60%
	Hyperparameters															
	-1	96	67%	72%	66%	62%	67%	72%	100%	62%	99%	48%	91%	72%	69%	69%
	0	89	61%	44%	63%	56%	60%	59%	100%	44%	65%	56%	77%	48%	60%	59%
	1	729	79%	78%	99%	75%	79%	78%	99%	64%	79%	78%	79%	78%	21%	23%
	2	1982	65%	62%	100%	61%	66%	63%	100%	54%	66%	63%	69%	63%	63%	62%
	3	252	65%	63%	69%	59%	68%	59%	99%	58%	63%	61%	75%	59%	48%	42%
	4	832	70%	64%	98%	62%	87%	66%	98%	60%	70%	58%	72%	66%	63%	61%
	5	132	62%	60%	72%	60%	87%	65%	100%	68%	71%	68%	88%	73%	64%	58%
	6	40	86%	67%	86%	67%	86%	58%	100%	67%	86%	67%	86%	67%	86%	67%
	7	14	78%	80%	67%	80%	89%	60%	100%	80%	89%	40%	89%	60%	89%	60%
	8	20	100%	83%	100%	83%	93%	100%	100%	67%	71%	83%	100%	100%	64%	83%
	9	14	67%	80%	67%	20%	78%	60%	89%	60%	67%	80%	89%	60%	78%	40%
	All Clusters	4200	67%	66%	75%	62%	69%	67%	99%	55%	68%	66%	71%	67%	64%	65%
2015	**Default Parameters**															
	-1	54	100%	65%	100%	59%	78%	71%	92%	53%	70%	65%	95%	71%	70%	71%
	0	2948	100%	71%	100%	70%	73%	70%	76%	77%	69%	70%	88%	77%	66%	62%
	1	626	99%	73%	99%	71%	74%	78%	78%	76%	65%	64%	87%	76%	71%	71%
	2	514	100%	75%	100%	72%	76%	77%	77%	74%	74%	71%	89%	80%	74%	75%
	3	12	100%	75%	100%	75%	88%	75%	100%	75%	75%	75%	100%	75%	88%	75%
	4	174	100%	51%	100%	60%	60%	40%	66%	68%	63%	72%	79%	72%	64%	79%

	5	75	100%	52%	100%	65%	77%	61%	87%	83%	63%	48%	100%	65%	63%	52%
	6	75	100%	70%	100%	70%	75%	78%	77%	65%	65%	65%	96%	65%	67%	74%
	7	19	100%	33%	100%	33%	92%	33%	92%	33%	69%	50%	100%	17%	69%	67%
	8	101	93%	74%	93%	81%	86%	74%	87%	74%	83%	77%	90%	77%	23%	26%
	All Clusters	4598	100%	69%	100%	72%	74%	73%	73%	75%	67%	68%	86%	77%	65%	65%
	Hyperparameters															
	-1	54	65%	65%	100%	59%	59%	59%	100%	65%	92%	59%	95%	71%	70%	71%
	0	2948	74%	70%	80%	71%	84%	76%	100%	71%	80%	74%	86%	77%	68%	67%
	1	626	73%	77%	81%	73%	78%	76%	99%	72%	77%	76%	82%	73%	71%	71%
	2	514	75%	74%	100%	72%	84%	81%	100%	71%	78%	77%	85%	81%	74%	75%
	3	12	75%	75%	75%	75%	100%	75%	100%	75%	100%	75%	100%	75%	88%	75%
	4	174	74%	60%	70%	62%	69%	74%	100%	58%	64%	77%	76%	72%	64%	79%
	5	75	65%	65%	100%	65%	79%	57%	100%	61%	98%	70%	83%	61%	65%	43%
	6	75	100%	65%	75%	70%	67%	78%	100%	65%	94%	65%	92%	70%	65%	65%
	7	19	69%	50%	69%	50%	69%	50%	100%	33%	69%	50%	85%	33%	69%	50%
	8	101	86%	74%	84%	77%	89%	74%	87%	65%	83%	77%	86%	77%	83%	77%
	All Clusters	4598	75%	72%	76%	72%	78%	76%	100%	71%	77%	72%	83%	77%	66%	67%
2016	**Default Parameters**															
	-1	100	100%	47%	100%	50%	76%	47%	77%	53%	59%	50%	94%	50%	51%	60%
	0	4218	100%	61%	100%	63%	65%	62%	66%	66%	62%	65%	81%	68%	55%	56%
	1	833	100%	64%	100%	68%	58%	60%	69%	68%	66%	66%	87%	72%	62%	64%
	2	1667	97%	56%	97%	56%	62%	64%	64%	62%	55%	53%	76%	64%	58%	65%
	3	299	99%	50%	99%	48%	61%	50%	67%	59%	60%	64%	81%	48%	59%	49%
	4	64	100%	55%	100%	50%	75%	30%	77%	35%	57%	60%	91%	55%	61%	55%
	5	36	100%	45%	100%	82%	84%	73%	84%	82%	68%	64%	96%	64%	72%	64%
	All Clusters	7217	99%	61%	99%	60%	63%	62%	64%	63%	62%	60%	78%	67%	40%	42%
	Hyperparameters															
	-1	100	54%	47%	63%	53%	70%	47%	100%	50%	91%	47%	80%	53%	59%	53%
	0	4218	68%	65%	70%	64%	69%	67%	100%	64%	75%	65%	78%	68%	62%	65%
	1	833	67%	66%	76%	69%	79%	69%	100%	61%	83%	69%	83%	72%	65%	66%
	2	1667	65%	60%	97%	57%	71%	61%	97%	51%	66%	61%	71%	63%	58%	65%
	3	299	64%	58%	99%	51%	65%	52%	99%	49%	64%	50%	77%	48%	59%	49%
	4	64	95%	45%	100%	55%	57%	60%	100%	40%	75%	55%	80%	45%	57%	70%
	5	36	84%	64%	80%	73%	88%	82%	100%	64%	88%	64%	88%	73%	72%	64%
	All Clusters	7217	67%	62%	70%	62%	71%	64%	99%	60%	73%	64%	75%	68%	62%	60%
2017	**Default Parameters**															
	-1	129	100%	69%	100%	59%	73%	62%	84%	69%	100%	67%	94%	69%	66%	51%
	0	3682	100%	62%	100%	63%	63%	62%	67%	65%	58%	54%	83%	69%	42%	46%
	1	480	99%	52%	99%	54%	60%	53%	64%	62%	56%	54%	81%	60%	45%	50%

	2	631	100%	63%	100%	69%	69%	71%	72%	75%	67%	72%	90%	75%	65%	74%
	3	171	99%	60%	99%	67%	71%	71%	73%	67%	62%	62%	85%	69%	66%	71%
	4	20	100%	50%	100%	50%	79%	83%	100%	67%	64%	50%	93%	67%	64%	100%
	5	22	100%	43%	100%	57%	80%	29%	100%	43%	60%	71%	93%	43%	47%	14%
	6	11	100%	25%	100%	25%	86%	0%	100%	25%	86%	0%	100%	25%	100%	25%
	7	11	100%	50%	100%	100%	86%	100%	100%	50%	86%	100%	100%	50%	86%	25%
	8	14	100%	60%	100%	40%	100%	40%	100%	60%	56%	40%	100%	60%	100%	40%
	All Clusters	5171	100%	61%	100%	62%	62%	61%	67%	64%	57%	56%	81%	69%	45%	45%
	Hyperparameters															
	-1	129	93%	56%	100%	64%	78%	59%	100%	62%	100%	64%	89%	72%	66%	51%
	0	3682	67%	62%	100%	63%	73%	67%	100%	61%	83%	63%	80%	69%	58%	55%
	1	480	59%	50%	70%	56%	65%	56%	99%	50%	76%	52%	71%	59%	45%	49%
	2	631	62%	66%	80%	70%	89%	72%	100%	63%	79%	74%	86%	77%	67%	72%
	3	171	75%	71%	68%	71%	78%	71%	99%	58%	71%	69%	84%	71%	66%	71%
	4	20	93%	67%	79%	100%	86%	67%	100%	50%	86%	67%	93%	67%	71%	83%
	5	22	60%	71%	67%	57%	80%	71%	100%	57%	80%	57%	87%	43%	47%	14%
	6	11	100%	0%	86%	0%	100%	25%	100%	25%	100%	25%	100%	25%	100%	25%
	7	11	86%	100%	86%	100%	100%	100%	100%	100%	86%	100%	100%	50%	86%	25%
	8	14	100%	40%	100%	40%	100%	40%	100%	60%	100%	40%	100%	60%	100%	40%
	All Clusters	5171	66%	63%	100%	63%	77%	68%	100%	61%	73%	64%	77%	68%	56%	56%
2018	**Default Parameters**															
	-1	48	100%	53%	100%	80%	85%	80%	94%	60%	67%	67%	94%	67%	76%	60%
	0	1129	100%	64%	100%	65%	62%	56%	70%	66%	58%	55%	84%	68%	57%	52%
	1	5763	100%	62%	100%	64%	69%	69%	69%	70%	57%	59%	80%	72%	57%	58%
	2	1056	100%	63%	100%	68%	71%	74%	74%	77%	62%	68%	86%	78%	63%	68%
	3	393	99%	67%	99%	67%	63%	57%	75%	73%	72%	72%	83%	73%	71%	70%
	4	21	100%	57%	100%	71%	93%	86%	100%	86%	79%	86%	100%	71%	86%	71%
	All Clusters	8410	100%	62%	100%	65%	67%	66%	70%	68%	60%	58%	80%	71%	59%	58%
	Hyperparameters															
	-1	48	100%	80%	76%	73%	88%	67%	100%	67%	94%	73%	91%	73%	73%	60%
	0	1129	71%	63%	72%	64%	81%	65%	100%	60%	74%	65%	81%	68%	58%	56%
	1	5763	70%	65%	100%	65%	72%	68%	100%	63%	76%	68%	78%	72%	57%	58%
	2	1056	78%	69%	100%	67%	74%	73%	100%	64%	81%	74%	83%	78%	63%	67%
	3	393	72%	72%	77%	69%	71%	72%	99%	68%	75%	72%	80%	71%	72%	72%
	4	21	100%	86%	86%	71%	100%	71%	100%	43%	93%	71%	93%	71%	79%	86%
	All Clusters	8410	73%	67%	100%	66%	73%	69%	100%	63%	77%	68%	76%	72%	59%	58%

Appendix B: Supplemental materials to Chapter 4 – Objective 3

Annual Pearson's correlation coefficient between number of confirmed cases and related significant factors.

Initial layer: 2012 **Final layer:** 2013
Spatial variables: Elevation, cases, LULC, Temperature, wind speed, precipitation, population, humidity

Table B.1. 2012 Pearson's correlation coefficient

	Cases	Humidity	Elevation	Wind Speed	Precipitation	Temperature	Population	LULC
Cases	-	0.21	-0.13	0.09	-0.05	0.07	0.48	-0.53
Humidity		-	-0.54	-0.14	-0.15	0.09	0.17	-0.38
Elevation			-	0.46	0.15	0.11	-0.12	0.19
Wind Speed				-	0.17	0.71	-0.01	-0.12
Precipitation					-	0.11	-0.04	0.06
Temperature						-	0.02	-0.14
Population							-	-0.44
LULC								-

Initial layer: 2013 **Final layer:** 2014
Spatial variables: Elevation, cases, LULC, Temperature, wind speed, precipitation, population, humidity

Table B.2. 2013 Pearson's correlation coefficient

	Cases	Humidity	Elevation	Wind Speed	Precipitation	Temperature	Population	LULC
Cases	-	0.23	-0.05	0.18	-0.06	-0.03	0.46	-0.54
Humidity		-	-0.41	-0.01	-0.18	-0.01	0.21	-0.41
Elevation			-	0.37	0.03	0.05	-0.14	0.19
Wind Speed				-	-0.06	0.24	0.04	-0.19
Precipitation					-	-0.02	-0.04	0.03
Temperature						-	-0.03	0.05
Population							-	-0.49
LULC								-

Initial layer: 2014 **Final layer:** 2015
Spatial variables: Elevation, cases, LULC, Temperature, wind speed, precipitation, population, humidity

Table B.3. 2014 Pearson's correlation coefficient

	Cases	Humidity	Elevation	Wind Speed	Precipitation	Temperature	Population	LULC
Cases	-	0.34	-0.15	0.07	-0.04	0.01	0.52	-0.55
Humidity		-	-0.41	-0.02	-0.01	-0.07	0.23	-0.41
Elevation			-	0.29	0.03	0.31	-0.15	0.19
Wind Speed				-	0.09	0.59	0.02	-0.15
Precipitation					-	0.06	0.002	-0.11
Temperature						-	0.01	-0.03
Population							-	-0.53
LULC								-

Initial layer: 2015 **Final layer:** 2016
Spatial variables: Elevation, cases, LULC, Temperature, wind speed, precipitation, population, humidity

Table B.4. 2015 Pearson's correlation coefficient

	Cases	Humidity	Elevation	Wind Speed	Precipitation	Temperature	Population	LULC
Cases	-	0.32	-0.24	0.06	-0.20	-0.0007	0.47	-0.43
Humidity		-	-0.41	0.09	-0.27	-0.07	0.24	-0.41
Elevation			-	0.26	0.09	0.31	-0.16	0.19
Wind Speed				-	0.13	0.59	0.04	-0.20
Precipitation					-	0.05	-0.07	0.09
Temperature						-	0.01	-0.03
Population							-	-0.54
LULC								-

Initial layer: 2016 **Final layer:** 2017
Spatial variables: Elevation, cases, LULC, Temperature, wind speed, precipitation, population, humidity

Table B.5. 2016 Pearson's correlation coefficient

	Cases	Humidity	Elevation	Wind Speed	Precipitation	Temperature	Population	LULC
Cases	-	0.20	-0.15	0.11	-0.07	-0.05	0.47	-0.41
Humidity		-	-0.42	0.03	-0.22	-0.07	0.24	-0.41
Elevation			-	0.38	0.07	0.03	-0.16	0.19
Wind Speed				-	-0.07	0.19	0.09	-0.24
Precipitation					-	0.02	-0.06	0.10
Temperature						-	-0.04	0.07
Population							-	-0.54
LULC								-