

**An Intelligent Prediction Model for Infectious Diseases' Outbreaks:
An Ensemble of Machine Learning, Big Climate Data and
Indigenous Knowledge**

MAINGOSINATHI PAULINA PHOOBANE

Thesis submitted in fulfilment of the requirements for the Degree

DOCTOR OF PHILOSOPHY IN INFORMATION TECHNOLOGY

in the

Department of Electrical, Electronic and Computer Engineering
Faculty of Engineering, Built Environment and Information Technology

at the

Central University of Technology, Free State

Promoter: Prof EM Masinde (PhD. Computer Science)

Co-promoter: Prof OJ Botai (PhD. Meteorology)

Co-promoter: Prof T Mabhaudhi (PhD. Crop Science)

DECEMBER 2024

Copyright Notice

This doctoral thesis is to be used for only academic or non-commercial research purposes. The information contained in this thesis is to be published with acknowledgement of the source. This doctoral thesis is published by the Central University of Technology, Free State, in terms of a non-exclusive licence granted to CUT by the author.

Disclaimer

The presentation of this thesis contains colour images meant to present the results visually in a more understandable form. While printing or viewing the thesis in greyscale (black and white) is possible, it is recommended that for clarity, the thesis is viewed (or printed) in full-colour format

Dedication

This thesis is dedicated to my late grandmother, Mafitisa Alina Mosuhli and all my family members and friends for their kind support.

Declaration

This research, as presented in this thesis, is my original work and has not been presented for any other university award. Knowledge derived from other sources has been clearly indicated, with acknowledgement and reference to the literature.

This study was conducted and completed under the guidance of Professor Muthoni Masinde, Department of Information Technology at Central University of Technology, Free State, South Africa, co-supervised by Professor Tafadzwanashe Mabhaudhi, School of Agricultural, Earth and Environmental Sciences at the Kwa-Zulu Natal University and Professor Botai Joel, Department of School of Agricultural, Earth and Environmental Sciences at Kwa-Zulu Natal University, Free State, South Africa.

Maingosinathi Paulina Phoobane

Signature: _____

Date: 03 December 2024

In our capacity as the supervisors of this thesis, we certify that the above statements are true to the best of our knowledge.

Professor Muthoni Masinde

Signature: _____

Date: 04 December 2024

Professor Tafadzwanashe Mabhaudhi

Signature: _____

Date: 04 December 2024

Professor Botai Joel

Signature: _____

Date: 04 December 2024

Preface and Acknowledgements

This PhD thesis represents the culmination of years of dedication, research and the invaluable support of many individuals and organizations. I am profoundly grateful to all who have contributed to this journey.

First and foremost, I would like to extend my heartfelt thanks to my supervisor, Prof. Muthoni Masinde, whose guidance, wisdom and encouragement have been a constant source of inspiration throughout this journey. I also deeply appreciate the support and insights provided by my co-supervisors, whose expertise has greatly enhanced the quality of this work.

I am equally indebted to the Vhembe community for their invaluable contribution to Indigenous Knowledge data. Their willingness to share their knowledge and experiences has enriched this research and provided a unique perspective that has been crucial to the success of this project.

I would like to express my sincere gratitude to the South African Weather Service for their generous support in providing the essential meteorological data required for this research. Their assistance was instrumental in enabling the analysis and development of the models presented in this study.

My deepest gratitude goes to my mom, 'Manako Phoobane; my dad, Thabiso Phoobane; my aunt, Sr Alphoncina Mosuhli; my sons, Teboho and Tlotliso Makoetlane; and, to all my family and friends. Your unwavering support and belief in my abilities have sustained me through the challenges of this journey. Your encouragement has been a source of strength, and I am forever grateful for your love and patience.

Finally, I would like to give thanks to God for granting me the perseverance, strength and wisdom to complete this work. This thesis would not have been possible without the grace and blessings I have received.

To all who have supported me in this endeavour, I offer my sincere thanks and appreciation. This achievement is as much yours as it is mine.

Abstract

Climate change is real. It affects all spheres of fauna and flora, remarkably increasing human vulnerability to infectious diseases. Infectious diseases are the major causes of death in low-income countries, with Sub-Saharan Africa accounting for a larger proportion of these fatalities. One of the challenges in fighting infectious diseases in developing countries like South Africa is the lack of effective surveillance systems for predicting and monitoring infectious diseases. It is essential to provide an effective, relevant early warning system (EWS) for infectious disease outbreaks to mitigate impact through vaccination and other interventions.

Although several studies have developed early warning systems or models to predict and/or monitor infectious disease outbreaks such as malaria, these studies are predominately based only on one knowledge system, primarily focused on scientific approaches and do not account for an African context. The integration of these current approaches with indigenous knowledge (IK) systems could enhance current early warning systems and make them more effective, relevant and understandable to the communities the EWS are intended to serve. The objective of this research study is to develop an intelligent people-centred model for predicting infectious disease outbreaks using IK, big climate data and historical malaria incidence in the Vhembe district of South Africa. Vhembe is a district in the province of Limpopo that is identified as prone to malaria outbreaks and other climate-driven shocks.

The lens that guided this research study is the EWS framework termed Malaria Outbreak Early Warning System (MOEWS). MOEWS encompasses the collection and analysis of weather and IK data to predict malaria outbreaks. Malaria outbreak risk knowledge encompasses collecting, understanding and integrating historical weather, malaria and IK data with drought indices. Machine learning (ML) models were used to predict malaria outbreaks using the collected data.

Both classification and regression ML algorithms were explored to predict malaria using structured data; weather, malaria and drought indices. Data pre-processing, such as normalization of data, handling outliers, oversampling and feature selections, was performed before modelling the dataset. ML models, developed in Jupiter Notebook using Python, were trained and tested using historical weather, drought indices and malaria data. The Multilayer Perceptron classification showed an optimal

performance. MLP model demonstrated an accuracy of 93%, a precision of 95%, a perfect recall of 100%, and an F1 score of 96%, indicating a robust predictive capability. As a result, the MLP model was selected for malaria prediction. The model was subsequently deployed to make malaria predictions using real-time data on weather conditions and reported malaria incidences. A lag of one to two months was applied between the predictors and the malaria outbreak.

The results suggest the MLP's robustness in identifying patterns and predicting malaria outbreaks effectively. The deployment of the MLP model with real-time data on weather conditions and reported malaria incidences highlights its practical application in forecasting malaria outbreaks. This implies that the model has the potential to serve as a reliable tool in early warning systems, enabling proactive interventions and informed decision-making for malaria control and prevention.

For the malaria outbreak forecast using indigenous knowledge, IK indicators for malaria forecast collected from the local people in Vhembe were formatted and represented as concepts. The relationship/causal effects between the concepts and with malaria were determined and formally represented as adjacency matrices. Adjacency matrices were used to develop fuzzy cognitive maps (FCM) for each of the four seasons of the year, illustrating the concepts and their causal relationships. The importance of each IK indicator within each seasonal FCM was established by analysing the density, in-degree, out-degree and centrality measures. The findings highlight that in Autumn, "autumn heavy rainfalls" and "dirty water in containers/small pools" are major predictors of malaria outbreak while "summer heavy rainfalls," "dirty water in containers" and "summer temperature" are key indicators in summer. In contrast, the main predictors of malaria in winter are 'Fig "Muhuyu" trees not shedding leaves' and 'sight of insects/locusts' while spring rainfalls and "'Mofafa grass" having many ticks' were found important for malaria prediction in spring. These insights imply that malaria forecasting systems can benefit from incorporating seasonally relevant IK indicators to improve accuracy and contextual relevance.

The autumn FCM model was utilised to forecast seasonal malaria outbreaks. This was done using real-time indigenous knowledge indicators reported by selected local experts via the mobile MOEWS App. Based on the malaria IK indicators reported by

the IK expert for autumn, the malaria outbreak concept had a value of 0.7856 indicating that a low malaria incidence ranging from 400 to 800 cases could be expected.

Malaria alerts generated from IK and machine learning-based forecasts were disseminated through the MOEWS mobile App, social media platforms and a web portal. The MOEWS App plays a crucial role in this process, providing a direct and immediate connection between the forecast and the community. The alerts are intended to aid policymakers in developing mitigation strategies and informing decision-making processes related to malaria. This ensures that the policymakers are well-informed and can respond swiftly to potential outbreaks.

List of Acronyms

AI	Artificial Intelligence
ANN	Artificial Neuron Networks
AWRI	Available Water Resource Index
CART	Classification and Regression Trees
CRISP	Cross-Industry Standard Process for Data Mining
CUT	Central University of Technology
DEP	Deviation of Effective Precipitation
DL	Deep Learning
DT	Decision Trees
EDI	Effective Drought Indices
EP	Effective Precipitation
EWS	Early Warning System
FCM	Fuzzy Cognitive Maps
GIS	Geographical Information System
GPS	Global Positioning System
HDFS	Hadoop's Distributed File System
ICMEWS	Integrated Climate-driven Multi-hazard Early Warning System for South Africa
ICT	Information and Communication Technology
IK	Indigenous Knowledge
IKI	Indigenous Knowledge Indicators
IKS	Indigenous Knowledge System
KNN	K-Nearest Neighbours
LSTM	Long Short-Term Memory
MAE	Mean Absolute Error
MDP	Markov Decision Process
MEP	Mean Effective Precipitation
ML	Machine Learning

MLP	Multilayer Perceptron
MOEWS	Malaria Outbreak Early Warning System
MSE	Mean Square Error
PDS	Percentage Degree Severity
PPL	Python Programming Language
RF	Random Forest
RL	Reinforcement Learning
RMSE	Root Mean Square Error
RST	Rough Set Theory
SAWS	South Africa Weather Service
SL	Supervised Learning
SOP	Self-Organising Map
SPSS	Statistical Package for Social Sciences
SSA	Sub-Saharan Africa
SVM	Support Vector Machine
UL	Unsupervised Learning
UNISDR	United Nations International Strategy on Disaster Reduction

Table of Contents

Contents

Copyright Notice	i
Disclaimer.....	ii
Dedication	iii
Declaration	iv
Preface and Acknowledgements.....	v
Abstract	vi
List of Acronyms.....	ix
Table of Contents	xi
List of Figures	xvi
List of Tables.....	xix
List of Publications	xxi
1. Chapter One: Introduction and Background Information	1
1.1 Introduction	1
1.2 Problem Statement	3
1.3 Motivation and Justification.....	4
1.4 Research Questions.....	5
1.5 Research Aim and Objectives.....	6
1.6 Research Hypothesis	6
1.7 The Solution Approach.....	7
1.8 Significance and Contribution of the Study	7
1.9 Evaluation Criteria.....	8
1.10 Scope and Limitation.....	9
1.11 Thesis Structure	9
2. Chapter Two: Literature Review	11
2.1 Introduction	11
2.2 Bibliometric Analysis of Infectious Disease Prediction in Africa	11
2.2.1 Predicting Infectious Disease Outbreaks Research Trends	13
2.2.2 Network Analysis.....	15
2.3 Infectious Disease Prediction.....	19

2.3.1	Malaria	21
2.3.2	Infectious Disease Prediction and Information Communication and Technology - Where is the Link?.....	24
2.4	AI Techniques in Predicting Infectious Diseases Associated with Climate Change..	27
2.4.1	Artificial Neural Networks and Deep Learning	33
2.4.2	Support Vector Machine	36
2.4.3	Decision Trees	37
2.4.4	Random Forest	38
2.4.5	Fuzzy Cognitive Map	39
2.5	Big Climate Data	41
2.5.1	Big Data	41
2.5.2	Big Climate Data	43
2.5.3	Processing Big Climate Data	43
2.6	Indigenous Knowledge in Infectious Disease Outbreak Prediction	48
2.6.1	Definition of Indigenous Knowledge.....	48
2.6.2	Indigenous Knowledge Indicators for Malaria Prediction.....	49
2.6.3	Infectious Disease Prediction Using IK vs Modern Science.....	51
2.6.4	Differences and Similarities between IK and Modern Science	52
2.6.5	Indigenous Knowledge and Modern Science Integration	54
2.7	Early Warning Systems	54
2.7.1	Early Warning System for Infectious Diseases	55
2.7.2	People-Centred Early Warning Systems.....	56
2.8	Conclusion.....	56
3	Chapter Three: Research Design and Methodology	58
3.1	Introduction	58
3.2	Malaria Outbreak Early Warning System (MOEWS) as a Framework.....	58
3.2.1	General Elements of an Early Warning System.....	59
3.2.2	MOEWS Framework Architecture	61
3.2.3	Features of Malaria Outbreak Early Warning System	64
3.2.4	Summary of the MOEWS Framework.....	65
3.3	Research Design and Methodology	66
3.3.1	Research Design	68

3.3.2	Study Area	70
3.3.3	Ethical Consideration	71
3.3.4	Target Population	72
3.3.5	Sampling Techniques.....	72
3.3.6	Data Collection	72
3.3.7	Data Types.....	76
3.3.8	Data Analysis and Interpretation	77
3.3.9	Reliability and Validity.....	80
3.3.10	Summary Research Design and Methodology.....	81
3.4	Chapter Conclusion	81
4	Chapter Four: Infectious Disease Outbreak Risk Knowledge.....	83
4.1	Introduction	83
4.2	Data Source 1: Indigenous Knowledge	84
4.2.1	Sample Data	84
4.2.2	Data Presentation, Pre-Analysis, and Discussion.....	85
4.2.3	IK Indicators.....	101
4.3	Data source 2: Conventional Weather Station Data Collection	106
4.3.1	Temperature.....	106
4.3.2	Relative Humidity.....	107
4.3.3	Rainfall.....	108
4.3.4	Drought Indices.....	110
4.4	Data Source 3: Structured Health Data.....	113
4.5	Conclusion.....	114
5	Chapter Five: Predicting Malaria Outcome Using Indigenous Knowledge.....	116
5.1	Introduction	116
5.2	Fuzzy Cognitive Map Methodology	116
5.3	Identification and Analysis of Concepts for Malaria Prediction.....	117
5.3.1	Determining the Causal Effects Among Concepts.....	119
5.3.2	Building Weighted Adjacency Matrices	125
5.4	Representation of Concepts in Fuzzy Cognitive Maps	125
5.5	Representation of Causal Effects of Concepts Using Weighted Adjacency Matrices	125

5.6	Design and Building of Fuzzy Cognitive Maps.....	129
5.7	Inference	134
5.7.1	Representation of the input concepts vector	134
5.7.2	Representation of Causal Relationships between Concepts	135
5.7.3	Activation of New Concepts States and States of Equilibrium	135
5.8	Interpretation: The Likelihood of Malaria Outcomes	137
5.9	Conclusion.....	138
6.	Chapter Six: Malaria Prediction Using Weather and Malaria Data	140
6.1	Introduction	140
6.2	Dataset	140
6.3	Data Pre-processing	143
6.3.1	Data Cleaning	143
6.3.2	Merging Datasets of Different Time Series	143
6.3.3	Missing Values.....	143
6.3.4	Outliers.....	144
6.3.5	Data Normalisation	147
6.4	Feature Analysis	147
6.5	Regression Models Results	149
6.6	Classification Models	150
6.6.1	Converting Continuous Variables to Categorical Variables.....	151
6.6.2	Handling Imbalanced Data	153
6.6.3	Classification Models Results	155
6.7	Discussion of Results.....	160
6.8	Conclusion.....	161
7.	Chapter Seven: Malaria Outbreaks Communication and Dissemination.....	163
7.1	Introduction	163
7.2	System Overview and Use Case	163
7.3	Mobile Phone Applications Implementation	168
7.3.1	Application Installation	168
7.3.2	Application Functionalities.....	169
7.4	Web Portal	176
7.4.1	Text View of Malaria Predictions	176

7.4.2	Graphical View of Malaria Outbreak Predictions.....	177
7.5	Social Media	178
7.6	Conclusion	179
8.	Chapter Eight: Conclusion and Future Studies	180
8.1	Introduction	180
8.2	Evaluation of the Research Objectives.....	180
8.2.1	Climatic Variabilities, IK Indicators and Their Effects on Infectious Disease Outbreaks.....	180
8.2.2	Malaria Early Warning System Framework.....	181
8.2.3	Fuzzy Cognitive Maps for Malaria Outbreak Prediction	182
8.2.4	Machine Learning Model for Malaria Outbreak Prediction Using Weather Data	183
8.2.4	Building, Testing and Evaluating the System Prototype.....	185
8.3	Innovative Contributions of the Research Thesis	187
8.4	Recommendations and Future Directions	188
8.4.1	Enhancing MOEWS as a framework.....	188
8.4.2	Malaria Outbreak Risk Knowledge.....	188
8.4.3	Malaria Prediction.....	189
8.4.4	Malaria Dissemination and Communication.....	189
	List of References	190
	Appendix A: IK Indicators for Malaria Forecast in SSA.....	214
	Appendix B: Ethical Clearance	223
	Appendix C: Certificate of Completion of Research Ethics Module.....	224
	Appendix D: Questionnaire	225

List of Figures

Figure 2. 1: Distribution of annual publications on predicting infectious disease outbreaks research from 2011 to 2020 in Africa.	13
Figure 2.2: Countries in the top ten ranks in publications on predicting infectious disease outbreaks from 2011 to 2020.	15
Figure 2.3: Keyword co-occurrence network.....	16
Figure 2.4: Thematic map of emerging themes in predicting infectious diseases research in Africa.	18
Figure 2.5: Top 10 causes of death in low-income countries from 2000 to 2019 (World Health Organization, 2020b).....	20
Figure 2.6: ANN basic architecture(<i>Abiodun et al., 2018</i>).	34
Figure 2.7: A simplified version of SVM (Kufel et al., 2023).....	36
Figure 2.8: FCM structure (Papageorgiou et al., 2018).....	40
Figure 2.9: Sources of big data (Manogaran, Lopez and Chilamkurti, 2018).....	42
Figure 3.1: The components of an early warning system for infectious diseases.	60
Figure 3.2: Elements of MOEWS.	61
Figure 3.3: Malaria Outbreak Early Warning System Framework.....	62
Figure 3. 4: Detailed Malaria Early Warning System Architecture.....	64
Figure 3.5: Methodology flowchart.	67
Figure 3.6: Vhembe district map (Vhembe District Municipality, 2024)	70
Figure 4. 1: Participants distribution by migration history, age, and gender.	86
Figure 4. 2: Respondents' level of education by age and gender.	87
Figure 4. 3: Frequency of checking weather information by Age.	88
Figure 4. 4: The rating of the sources of weather information.....	89
Figure 4. 5: Weather-related alerts received by participants in Vhembe and the channels of communication.	90
Figure 4. 6: level of confidence in the accuracy of the sources of weather forecast information.	91
Figure 4.7: confidence in general use of IK according to gender and age.	92
Figure 4.8: Participants' confidence in general use of IK by migration history.....	95
Figure 4. 9: Confidence in IK use for disease prediction by gender and age.	96
Figure 4.10: Confidence in the IK use for farming according to migration history.	100
Figure 4.11: Categories of IKIs used by participants from the Vhembe Community.....	102
Figure 4.12: the monthly average temperature in Vhembe from 1998 to 2018.	107
Figure 4.13: Vhembe monthly average humidity from 1998 to 2018.	108
Figure 4.14: Climate and administrative districts of Limpopo.	109
Figure 4.15: daily rainfall from Weather District 35 in Vhembe from 1998 to 2018.	109
Figure 4.16: daily rainfall from Weather District 50 in Vhembe from 1998 to 2018.	110
Figure 4.17: EDI from rainfall recorded from 1998 to 2018 from District 35.	111
Figure 4.18: EDI from rainfall recorded from 1998 to 2018 from District 35.	111

Figure 4.19: AWRI from 1998 to 2018 from District 35.	112
Figure 4.20: AWRI from 1998 to 2018 from District 50.	112
Figure 4.21: Malaria incidences in Vhembe compared to the other four administrative districts in Limpopo from 1998 to 2018.....	113
Figure 4.22: malaria incidences in Vhembe according to different months of the year (from 1998 to 2018).....	114
Figure 5.1: Process of predicting malaria outbreaks using FCM.....	117
Figure 5.2: FCM for autumn.	129
Figure 5.3: FCM for Summer.	131
Figure 5.4: Winter FCM.	132
Figure 5.5: FCM for Spring.	133
Figure 5.6: Process of determining number of concepts and concepts values in FCM (Mwagha, 2016).	135
Figure 5.7: Graphical illustration of concept values through iterations of the inference process.	137
Figure 6.1: A sample of the dataset showing all the variables.	141
Figure 6.2: Methodology taken to build ML regression models to predict malaria incidences.	142
Figure 6.3: Boxplot indicating the presence of outliers in the dataset.....	145
Figure 6.4: Python code for identifying outliers and non-outliers in the dataset.	145
Figure 6.5: A view of some outliers in the dataset.	146
Figure 6.6: Addressing outliers using the median.	146
Figure 6.7: StandardScaler() function to scale the dataset.	147
Figure 6.8: The process of identifying an optimal classification model.....	151
Figure 6.9: dataset showing categorical.....	152
Figure 6. 11: Imbalanced dataset.....	153
Figure 6.12: Code snippet for oversampling data.....	154
Figure 6.13: Distribution of the malaria categories/classes after oversampling.	154
Figure 6.14: Comparing the accuracy score of the models.	155
Figure 6.15: Precision score of the models.....	156
Figure 6.16: Recall score achieved by the models.	156
Figure 6.17: F1 score of the models.....	157
Figure 6. 18: Graphical view predicted classes vs actual classes.	159
Figure 6. 19: Confusion matrix for MLP during the deployment.	159
Figure 7.1: MOEWS prototype architecture.	164
Figure 7.2: MOEWS use case diagram.	166
Figure 7.3: MOEWS Entity Relationship Diagram.	167
Figure 7.4: icon of MOWES mobile application.	169
Figure 7.5: MOEWS registration and login screen.	170
Figure 7.6: MOEWS main menu.....	171
Figure 7.7: MOEWS screens for reporting IK indicators.....	172

Figure 7.8: MOEWS' report extreme weather screen.	173
Figure 7.9: MOEWS view reported IK indicator screen.....	173
Figure 7.10: MOEWS view extreme weather events screen.....	174
Figure 7.11: MOEWS screens showing the previous predictions done in 2023 using weather patterns.	175
Figure 7.12: MOEWS screens showing the previous predictions done in Autumn 2023 using IK.	175
Figure 7. 13: Malaria prediction for January 2024.....	176
Figure 7.14: malaria prediction in January 2024.	177
Figure 7.15: graphical view of weather parameters.	177
Figure 7.16: IK indicators observed in Autumn 2023.....	178
Figure 7.17: Facebook page of MOEWS.....	178
Figure 8. 1: MLP performance during the test phase.	185
Figure 8. 2: MOEWS mobile application interface showing functionalities of the prototype.	186
Figure 8. 3: Predicted classes vs actual classes.....	187

List of Tables

Table 2.1: Severe outbreaks / first outbreak detection of infectious diseases in Africa from 2010 to 2020.	14
Table 2.2: Estimated number of malaria deaths by WHO region (<i>World Health Organization, 2022</i>)	22
Table 2.3: AI algorithms in infectious disease prediction.	28
Table 2. 4: Confusion matrix for binary classification problem (Vujovic, 2021).	33
Table 2. 5: The complementary aspect of IK and Modern Science.....	52
Table 2.6: The differences and similarities between LK and modern science forecasts on disease outbreaks.	53
Table 4. 1: EDI values and their corresponding interpretation.	78
Table 4. 2 Participants' distribution by Mobile ownership and Usage.	87
Table 4.3: Use of IK by the Vhembe community.	91
Table 4. 4: Contingent table for confidence in general use of IK according to gender.	92
Table 4.5: Contingency table for the confidence in general IK use according to age.	94
Table 4.6: Cramer's V and interpretation (Akoglu, 2018).	94
Table 4. 7: Contingent table for the general use of IK by migration history.	95
Table 4. 8: Contingent table for gender and confidence in IK use for disease prediction.	97
Table 4.9: Contingency table of confidence in IK use for disease prediction by age.	97
Table 4.10: Contingent table for the confidence in the use of IK for disease prediction by migration history.	98
Table 4. 11: Contingency table for use of IK for farming by migration history.	100
Table 4.12: Summary of the confidence participants from the Vhembe community have in IK use according to gender, age, and migration history.	101
Table 4.13: Weather Seasons in Vhembe District.	102
Table 4. 14: Malaria IKIs for the Vhembe community.....	104
Table 5.1: Malaria outbreak prediction concepts.	118
Table 5.2: Causal effects of concepts (IKIs) on malaria outbreak during summer.	119
Table 5.3: Causal effects of concepts (IKIs) on malaria incidences in autumn.	120
Table 5.4: Causal effects of concepts (IKIs) on malaria outbreak in winter.	120
Table 5.5: Causal effects of concepts (IKIs) on malaria outbreak in spring.....	120
Table 5.6: Summary of mode and mean of concepts (IKIs) causal effects on malaria according to seasons.	121
Table 5.7: Causal effects between IKIs concepts in Autumn.....	122
Table 5.8: Causal effects between IKIs concepts in Winter.....	123
Table 5.9: Causal effects between IKIs concepts in Spring.	123
Table 5.10: Weighted adjacency matrix for Summer.	127
Table 5.11: Weighted adjacency matrix for Autumn.	128
Table 5.12: Weighted adjacency matrix for Winter.....	128
Table 5.13: Weighted adjacency matrix for Spring.....	128

Table 5.14: Analysis of the importance of nodes/concepts in the FCM for Autumn.....	130
Table 5.15: Analysis of the importance of nodes/concepts in summer FCM.	131
Table 5.16: Analysis of the importance of nodes/concepts in Winter FCM.	133
Table 5.17: Analysis of the importance of nodes/concepts in Spring FCM.	133
Table 5.18: Concept values from the initial to the final iteration of the inference process. .	136
Table 5.19: Linguistic Variables and Corresponding Malaria Incidence Ranges Based on PDS.	138
Table 6. 1: Performance of interpolation techniques for patching missing data points.....	144
Table 6.2: Mutual information for features in the prediction of malaria in Vhembe District.	148
Table 6.3: Mutual Information after the optimum lag times.....	148
Table 6.4: Performance of the developed models.....	149
Table 6. 5:Malaria classes and their corresponding malaria incidences brackets with linguistic representation.....	153
Table 6.6: Weather parameters and drought indices used to deploy MLP model for March to October 2023 malaria prediction.....	158
Table 6.7: Malaria incidences reported in Vhembe in 2023.	158
Table 6. 8: Predicted malaria classes vs actual malaria classes.	158
Table 8.1: IK Indicators critical in malaria outbreak prediction in Vhembe.....	182

List of Publications

- Phoobane, P., Mabhaudhi, T., and Botai, J. 2023. Predicting Malaria Outbreak Using Indigenous Knowledge and Fuzzy Cognitive Maps: A Case Study of Vhembe District in South Africa. In International Conference on Emerging Technologies for Developing Countries (pp. 145-164). Cham: Springer Nature Switzerland https://doi.org/10.1007/978-3-031-63999-9_9
- Phoobane, P, and Masinde, M. 2023. Investigating the adoption of indigenous knowledge in mitigating climate-linked challenges: a case study of Vhembe District in South Africa. International Journal of Research in Business and Social Science (2147-4478), 12(7), pp.394-404. <https://doi.org/10.20525/ijrbs.v12i7.2728>
- Phoobane, P, Masinde, M. and Botai, J., 2022. Prediction Model for Malaria: An Ensemble of Machine Learning and Hydrological Drought Indices. In Proceedings of Sixth International Congress on Information and Communication Technology: ICICT 2021, London, Volume 3 (pp. 569-584). Springer Singapore. https://doi.org/10.1007/978-981-16-1781-2_51
- Phoobane, P., Masinde, M. and Mabhaudhi, T., 2022. "Predicting Infectious Diseases: A Bibliometric Review on Africa." *International Journal of Environmental Research and Public Health*. MDPI. <https://doi.org/10.3390/ijerph19031893>

1. Chapter One: Introduction and Background Information

1.1 Introduction

The increased incidences of infectious diseases, such as malaria, attributed to climate change pose a growing challenge to health systems across the globe (World Health Organization, 2023). These challenges are more pronounced in some regions, such as Sub-Saharan Africa (SSA), where current infectious disease surveillance systems are limited (Gaddy, 2020). The toll of infectious diseases is expected to increase significantly, especially in the context of climate change (Gaddy, 2020). Integrating IKSs with advanced Machine Learning (ML) and big climate data presents a novel approach that may enhance the robustness of early warning systems for infectious diseases. This study uses ML, big climate data, and indigenous knowledge (IK) to develop an early warning system for malaria. This approach addresses a critical gap in local health responses to infectious diseases such as malaria. By boosting IK with computational power, this study develops a more accurate and contextually relevant disease prediction system that may contribute to reducing outbreaks. Furthermore, the study may inform policy decisions intended to mitigate and manage malaria outbreaks.

Over the past three decades, the advancement of geographical information systems and satellite technology for earth observation has revolutionised the monitoring of weather, climate and environmental factors, including those influencing the emergence or re-emergence of infectious diseases (Ceccato *et al.*, 2018a; Tsatsaris *et al.*, 2021). Consequently, numerous infectious disease prediction tools have been developed to mitigate and manage outbreaks. However, many of these tools primarily focus on international or regional scales, often overlooking local-level disease coping mechanisms (Bansal *et al.*, 2016; Peiffer-Smadja *et al.*, 2020).

Various methods, spanning from traditional approaches to mathematical models and, more recently, ML techniques, have been employed to uncover hidden patterns in vast climate datasets. These methods aim to address numerous health challenges stemming from climate change and to assist in diagnosis, classification and risk forecasting. Few, however, predict future disease outbreaks (Sharma *et al.*, 2015; Masinde, 2020; Yadav, Perumal and Srinivas, 2020; Adamu, 2021). While most predictive

models rely on mathematical formulations, ML-based approaches remain underutilised in SSA despite their potential (Modu et al., 2017).

Predicting infectious disease outbreaks is inherently complex, particularly when incorporating climate data due to climatic systems' dynamic and intricate nature. This complexity underscores the necessity for effective prediction tools such as ML (Ceccato *et al.*, 2018a; Hassani, Huang and Silva, 2019). Furthermore, existing predictive tools predominantly adhere to a modern scientific approach, leveraging weather data, mathematical models and ML techniques while largely overlooking other knowledge systems, such as IK, which have demonstrated resilience to climate related shocks such as infectious disease outbreaks (Mashego, Maditsi and Bhuda, 2021). Therefore, integrating diverse knowledge systems into infectious disease prediction tools may enhance their robustness, leading to more effective and relevant early warning systems (EWS).

This research study focuses on the infectious disease domain with malaria prediction as a case study. Malaria is among the top 10 causes of death in low-income countries (World Health Organization, 2020a). Malaria claims the lives of over half a million people each year globally and Africa continues to bear the burden of morbidity (World Health Organization, 2022). Countries with a high burden of malaria generally have weak surveillance systems and cannot assess disease distribution trends. Consequently, many lives are lost (Amuakwa-Mensah, Marbuah and Mubanga, 2017; World Health Organization, 2022). Therefore, there is a dire need to build effective and relevant infectious disease predictive tools. One way to achieve this in Africa is to incorporate IK.

Factors such as climate and socio-economic parameters are often considered when predicting malaria, there being ample evidence in the literature linking malaria to these factors (Altizer *et al.*, 2013; Amuakwa-Mensah, Marbuah and Mubanga, 2017; Diouf *et al.*, 2022; Woyessa *et al.*, 2023). However, there are other factors, such as drought indices, that have not been explored in predicting malaria regardless of the remarkable success of their use in the prediction of climate phenomena such as drought (Kim, Byun and Choi, 2009; Masinde, Mwagha and Tadesse, 2018; Malik and Kumar, 2020a; Malik *et al.*, 2021). Many indices have been developed over the years, including the Effective Drought Index (EDI) and Available Water Resources Index (AWRI) (Byun and Wilhite,

1999) that express the actual amount of available water in the soil (Wilhite and Glantz, 1985). When it comes to malaria transmission, the latter could play a role (Phoobane, Masinde and Botai, 2022). Integrating data from various sources and employing advanced modelling techniques that borrow from different knowledge systems may enhance our understanding of malaria transmission dynamics and support effective prevention and control strategies such as early warning systems.

Given the technological constraints and low literacy rates in many rural and informal settlements in SSA, a bottom-up approach is imperative for implementing complex effective early warning systems for infectious disease outbreaks. This research study proposes an innovative EWS that uses two knowledge systems (modern science and IKS) integrating climate parameters, hydrological drought indices, malaria IK indicators and historical malaria outbreak data with various ML algorithms to predict malaria outbreaks.

1.2 Problem Statement

Substantial literature links the surge in recent decades in infectious disease outbreaks, such as malaria, to climate change (Diouf *et al.*, 2022; Woyessa *et al.*, 2023). With advancements in Artificial Intelligence (AI), big data, medical and climate data analysis has facilitated the prediction and early detection of infectious diseases associated with climate change (López *et al.*, 2022). Numerous factors have been considered in disease prediction, ranging from climate to socio-economic factors. Nevertheless, two drought indices, the AWRI and the EDI, remain underexplored as potential predictors for malaria outbreaks.

On the other hand, indigenous knowledge systems (IKS) have demonstrated resilience in combating infectious diseases linked to climatic variations (Macherera and Chimbari, 2016; Macherera, Chimbari and Mukaratirwa, 2017; Mafongoya *et al.*, 2017; Mashego, Maditsi and Bhuda, 2021). However, most existing Early Warning Systems (EWS) rely solely on modern science, primarily focusing on AI, to predict infectious disease outbreaks. While these EWSs often exhibit high prediction accuracy, they neglect other knowledge systems, such as IKS. Consequently, there is a need for studies that integrate modern computational techniques with IKS to develop more effective early warning systems that leverage the strengths of diverse knowledge systems. More intelligent, community-based early warning systems are needed to

predict infectious disease outbreaks related to weather in SSA. Such systems could inform policy decisions and raise awareness among healthcare professionals (e.g., doctors and nurses) and the public, allowing for preparation ahead of disease outbreaks. Infectious disease prevention and control measures, such as vaccinations, could be strategically planned and implemented in advance. Additionally, by involving the users for whom it is developed and incorporating their knowledge systems, such an approach could lead to a more effective, relevant and empowering EWS.

Overall, the following omissions and limitations were identified in the literature review, representing major weaknesses in current surveillance systems for malaria outbreak prediction in SSA:

- i) *Existing EWSs rely solely on one knowledge system.*
- ii) *Existing EWSs do not incorporate the coping mechanisms of the communities they are intended to serve, thus impeding their relevance and effectiveness.*
- iii) *Hydrological drought indices have not been incorporated into the current EWS despite the success of these indices in predicting phenomena such as drought.*

1.3 Motivation and Justification

Infectious diseases cause more deaths in low-income countries than non-communicable diseases and injuries. Malaria and other infectious diseases, such as diarrhoea and tuberculosis, are among the top ten causes of death in developing countries (World Health Organisation, 2020). Indeed, malaria remains a significant public health problem in many countries in Africa, particularly in SSA, where health resources are usually limited (World Health Organization, 2022). Over 50% of all malaria global deaths in 2022 occurred in four countries in SSA: Nigeria, the Democratic Republic of Congo, Mozambique and Uganda (World Health Organization, 2022). In South Africa, malaria is endemic in the low-altitude areas of the northern and eastern parts of South Africa along the border with Mozambique and Zimbabwe, with the transmission taking place mainly in Limpopo, Mpumalanga and Kwazulu-Natal provinces and occasionally in North-West province (Moonasar *et al.*, 2021). Limpopo accounts for more than half of the malaria burden in South Africa, with cases mostly occurring in the Vhembe district (National Department of Health South Africa, 2018).

There is a persistent threat posed by malaria, particularly in the SSA region, where healthcare resources are often limited. Despite ongoing efforts, traditional surveillance methods have proven insufficient in providing timely warnings for malaria outbreaks, leaving communities vulnerable to the devastating impacts of the disease. In response to this challenge, the proposed Malaria Outbreak Early Warning System (MOEWS) aims to fill this critical gap by harnessing a multidisciplinary approach. By amalgamating weather data, hydrological drought indices and scientific insights into malaria transmission patterns, MOEWS seeks to establish an advanced early warning system tailored to the specific needs of local communities. Through the deployment of MOEWS, the study aims to enhance preparedness and response capabilities, ultimately reducing the burden of malaria and safeguarding the health and well-being of vulnerable populations in the target regions.

Additionally, ample evidence exists of the link between climate variables and the surge in infectious disease outbreaks observed worldwide (Darkoh, Larbi and Lawer, 2017; Mussumeci and Codeço Coelho, 2020; Segun *et al.*, 2020; Zhao *et al.*, 2020; Hussain-Alkhateeb *et al.*, 2021; Ong *et al.*, 2023). With climate variability increasingly influencing disease transmission dynamics, there is a dire need for proactive measures to anticipate and mitigate the impacts of climate-related infectious diseases. On the other hand, like many other advancements, Africa has yet to keep pace in predicting infectious diseases regarding research output and using Fourth Industrial Revolution tools. Leveraging cutting-edge technologies such as ML and big data analysis presents a promising avenue for early detection and prediction of these diseases, offering valuable insights into emerging trends and enabling timely interventions. Furthermore, integrating IKS with AI and big data holds significant potential for enhancing resilience and fostering community-centred tools to combat infectious disease outbreaks effectively.

1.4 Research Questions

Building on the problem statement presented in the preceding section, this research study aims to address the following research questions:

1. What is the correlation between climatic variations and infectious disease outbreaks?

2. What machine learning algorithms are capable of efficiently and effectively predicting the correlation in question 1 above?
3. To what extent does the incorporation of local knowledge into a disease outbreak prediction model improve the model's intelligence, accuracy and relevance?

1.5 Research Aim and Objectives

The main objective of this research is to develop an intelligent system for predicting infectious disease outbreaks using selected indigenous knowledge, big climate data and historical outbreaks of selected infectious diseases in Southern Africa. A case study of the malaria outbreaks in Southern Africa is selected based on the pronounced incidences of malaria in Southern Africa and the availability and reliability of data.

Research Objectives

The main objective was accomplished through the following sub-objectives:

1. To analyse and understand climatic variations and their effects on outbreaks of infectious diseases.
2. To design a framework for an infectious disease early warning system that incorporates machine learning prediction algorithm(s) that model correlations between the selected infectious disease, indigenous knowledge and climatic variations.
3. To build, test and evaluate the effectiveness of a system prototype that implements the infectious disease early warning system for malaria outbreaks in Southern Africa.

1.6 Research Hypothesis

Integrating modern scientific methods (machine learning and weather parameters) with indigenous knowledge and drought indices in infectious disease prediction can create a more robust community-based early warning system that empowers local people.

1.7 The Solution Approach

The problem stated in Section 1.1 was addressed by developing a novel malaria outbreak EWS (MOEWS). Firstly, the MOEWS framework composed of three elements was developed: (1) Malaria Outbreak Risk Knowledge, (2) Malaria Outbreak Prediction, and (3) Malaria Outbreak Communication and Dissemination. Secondly, the three components were implemented. (1) Data on malaria outbreaks was collected and pre-processed. These data are from three domains: weather parameters, historical malaria incidences and the IK indicators associated with malaria outbreaks. (2) The data was then analysed using ML to forecast malaria outbreaks. (3) Malaria outbreak alerts were then disseminated to the relevant parties, such as the community, using a mobile phone app and web portal. Thirdly, the developed EWS, MOEWS, was evaluated using Vhembe as a case study. This thesis has documented the process undertaken in developing MOEWS, from its conceptualisation to development and evaluation.

1.8 Significance and Contribution of the Study

This research has significantly contributed to scientific knowledge through the novel early warning system for malaria outbreaks. An EWS system tailored for malaria outbreak predictions was developed. The developed EWS plays a crucial role in the early detection of malaria outbreaks, promptly alerting affected communities and all relevant parties. This rapid detection can significantly improve containment efforts and limit the spread of diseases. Moreover, MOEWS actively engages communities in malaria prediction and control efforts, empowering them with accessible information and encouraging proactive measures. The generated data on malaria outbreaks, climate factors, and IK indicators for malaria informs evidence-based decision-making and supports research and innovation in malaria monitoring and control, driving continuous improvement and innovation in combatting the disease. The finer detailed contributions of this research study are as follows:

- a) The research study has developed a novel early warning system framework for malaria prediction: Malaria Outbreak Early Warning System (MOEWS). MOEWS uses a hybrid technique that involves machine learning with IK, resulting in an EWS that is intelligent, effective, integrated, accessible and relevant to the people MOEWS is intended to serve. Multilayer Perceptron and

Fuzzy cognitive maps were used to make malaria outbreak forecasts. This is a significant contribution in the field of infectious diseases as MOEWS is a unique attempt to integrate IK and modern science to predict infectious diseases that are climate related.

- b) We used fuzzy cognitive maps to represent causal effects between IK indicators and malaria outbreaks. An understanding of the causal links between malaria IK indicators concepts was achieved using a case study to build up malaria outbreak knowledge. Statistical information was used to represent seasonal malaria outbreaks knowledge using fuzzy cognitive maps in the form of adjacency matrices. Incorporating IK into malaria outbreak prediction may enhance the effectiveness and relevance of our EWS. Involving the community in the development of our EWS ensures a people-centred EWS.
- c) This research study investigated the use of hydrological drought indices to forecast malaria outbreaks. The utilisation of hydrological drought indices, encompassing the mapping of the available water resource index and effective drought index to the ability of malaria-carrying vectors to thrive, represents a distinctive approach to predicting malaria outbreaks. Notably, extant literature predominantly focuses on mapping climate variables to malaria outbreak prediction, rendering this study's incorporation of drought indices a significant approach in infectious disease predictions. Moreover, the nature of the EDI and AWRI, characterised by their bottom-up methodology, allows them to capture intricate details about hydrological droughts, rendering them a robust predictor of malaria.

1.9 Evaluation Criteria

The research objectives stipulated above were tested against the outcome of the research. The EWS developed in this study used two approaches: modern science and IK. In modern science, ML algorithms are trained and tested. The best model was selected and used to make predictions. The predictions were compared to actual malaria incidences. Similarly, under IK, predictions of malaria outbreaks were made using the FCM and the results were compared with the actual data. A case study approach was adopted to verify and validate the developed EWS.

1.10 Scope and Limitation

The scope of this research is limited to predicting infectious diseases that are influenced by climatic changes in SSA, with a specific focus on malaria in the Vhembe district of the Limpopo province, South Africa. Due to time constraints, the study assumes it is impractical to include all IK indicators to predict malaria outbreaks. Therefore, only malaria IK indicators in the Vhembe district were considered in this case study.

1.11 Thesis Structure

The thesis is organised into eight chapters. The chapters were guided by the system framework, MOEWS. However, the second component of MOEWS was divided into chapters. The chapters were as follows:

- a) Chapter 1: This refers to the current chapter, which aims to provide an overview of the thesis.
- b) Chapter 2: This chapter offers an in-depth exploration of the existing literature that forms this study's foundation and conceptual framework. It highlights the theoretical underpinnings and technological advancements contributing to the primary achievements of this research. The bibliometric analysis is conducted to provide a birds eye view of the state of the art of the literature in the prediction of infectious diseases. It is followed by a systematic review whereby the impacts of malaria in SSA are highlighted. The use of ML and IK in infectious disease prediction was discussed, pointing out the gap in integrating these two knowledge systems to develop EWS for infectious diseases.
- c) Chapter 3: the chapter provides an overview of the research design and methodology involved in the study and the procedures that were followed to achieve research objectives and answer the research questions. This chapter also aims to present a comprehensive framework for the development of an Infectious disease EWS with a specific focus on malaria outbreak predictions (MOEWS). The three components are explained in detail, providing a blueprint for implementing a system that integrates IK and ML.
- d) Chapter 4: This chapter expounds upon implementing the first component of the system framework, which deals with collecting infectious disease outbreak risk knowledge. It analyses the heterogeneous data derived from three domains: 1. IK,

2. Conventional weather station data and 3. health data on historical malaria incidences to create malaria outbreak risk knowledge.
- e) Chapter 5: This chapter addresses the second component of MOEWS: prediction of malaria outbreaks. It describes a malaria prediction system using an IK approach. This approach uses fuzzy cognitive maps. The preliminary task was to understand the IK domain through statistical analysis of responses concerning IK for malaria outbreak prediction. Upon understanding the domain, the next step was to formulate the inputs, processing and outputs that define the mathematical logic that entails the malaria outbreak prediction model.
 - f) Chapter 6: This chapter, similar to Chapter 5, implements the second component of MOEWS but uses the modern science approach. It focuses on malaria prediction using a modern science approach. The chapter describes using weather data and drought indices to predict malaria outbreaks. It takes these data as an input and gives out monthly malaria forecasts as the output. Both regression and classification predictive models were explored to identify the optimal model to predict malaria outbreaks.
 - g) Chapter 7: Malaria Outbreak Communication and Dissemination. This chapter implements the third component of MOEWS. It discusses the two tools, a mobile application and a web portal, used to disseminate malaria outbreak alerts. The chapter further discusses the use of the social media platform Facebook, which allows users to engage with one another in discussions on various malaria-related topics and any other information shared on the Facebook page.
 - h) Chapter 8: This chapter concludes the thesis by presenting the results of MOEWS, followed by an evaluation of the research objectives. The chapter also summarises MOEWS' contributions to infectious disease predictions. Concluding remarks and recommendations for future work are discussed.

2. Chapter Two: Literature Review

2.1 Introduction

This chapter offers an in-depth exploration of the existing literature that forms this study's foundation and conceptual framework. It highlights the theoretical underpinnings and technological advancements that contribute to this research's primary achievements. It begins with a bibliometric analysis in Section 2.2, providing an overview of the current landscape in African infectious disease prediction research. Subsequent Sections 2.3 through 2.5 delve into the specifics of predicting infectious diseases, the application of AI techniques in the context of climate change and big climate data. The chapter also examines the use of indigenous knowledge in disease forecasting in Section 2.6 and discusses early warning systems in Section 2.7. The chapter concludes with a summary of the key findings, research gaps and insights in Section 2.8.

2.2 Bibliometric Analysis of Infectious Disease Prediction in Africa

At the onset of this research study, a bibliometric analysis was conducted to determine the state-of-the-art of the literature and to understand the research field. Bibliometric analysis is a quantitative method based on statistical and mathematical techniques that study the publication, dissemination and use of research (Pauna *et al.*, 2018). Even though the first effort of bibliometric analysis dates to the 1870s (Danesh and Mardani-Nejad, 2020), it has become a more powerful and necessary tool due to the proliferation and growth of scientific literature, which creates a challenge for researchers to review and process the mass of complex information being published in the literature (Fu *et al.*, 2023).

Bibliometric analysis was used in this research study to (1) identify the hot spot topics within infectious disease outbreak prediction in Africa, (2) analyse the relationship between infectious disease occurrence and the progression of publication, (3) determine the most contributing countries in this field and (4) most importantly, and due to the rich visualisations which bibliometric analysis offers, enable the researcher to present results in intuitive and comprehensible ways, providing a bird's-eye view and understanding of the available research knowledge in the prediction of infectious diseases in Africa (Vera-Polania *et al.*, 2014). The bibliometric analysis assisted in

conceptualising this research study and also resulted in a separate research output (Phoobane, Masinde and Mabhaudhi, 2022). Bibliometric analysis in any field is always aiming at a 'moving target' because of the continuous stream of publication. The analysis presented is based on the literature that existed on 1 October 2021 when the study was being conceptualised.

This research study focused on evolutionary trajectories, thematic analysis and emphasised keyword co-occurrence analysis. Keyword co-occurrence is a bibliometric network that allows for the visual exploration of the most important topics and their relatedness at a given time in a specific field of research (van Eck and Waltman, 2010). The aim of conducting a keyword co-occurrence network was to determine the direction researchers are taking in the field of infectious disease prediction and artificial intelligence (AI) or machine learning (ML) and big climate data. In keyword co-occurrence, the relatedness of the terms is based on the number of publications in which the terms occur together (van Eck and Waltman, 2017).

This research direction analysis process started with the advanced search in the Web of Science (WoS) core collection. The search was restricted to predicting infectious diseases using AI and big data, with Africa being the geographical boundary. No restrictions were set for language, document type or period of review (from January 1996 to 1 October 2021). Consequently, the search topic was set as follows: TS = ('infectious disease *' or 'COVID-19' or 'malaria' or 'Ebola' or 'plague' or 'Measles' or 'Yellow Fever virus' or 'monkeypox' or 'Zika Virus') AND TS = ('forecast *' or 'predict *' or 'machine learning' or 'artificial intelligence' or 'big data') AND TS = ('Africa').

The search resulted in a total of 1965 papers. However, the researcher screened the retrieved documents at the title, and abstract or full text to ensure data quality. Consequently, some documents were excluded for not meeting the inclusion criteria that the paper should be on human infectious disease prediction and conducted in Africa. The excluded documents were publications on predicting infectious diseases in animals and/or plants, documents on research not done in Africa, and duplicates or documents irrelevant to predicting infectious diseases. The filtering process resulted in 247 papers eligible for analysis. The retrieved data included articles, reviews, books and conference papers.

2.2.1 Predicting Infectious Disease Outbreaks Research Trends

Figure 2.1 below shows the evolutionary trajectory of the annual number of publications in predicting infectious disease outbreaks from 2011 to 2020 in Africa.

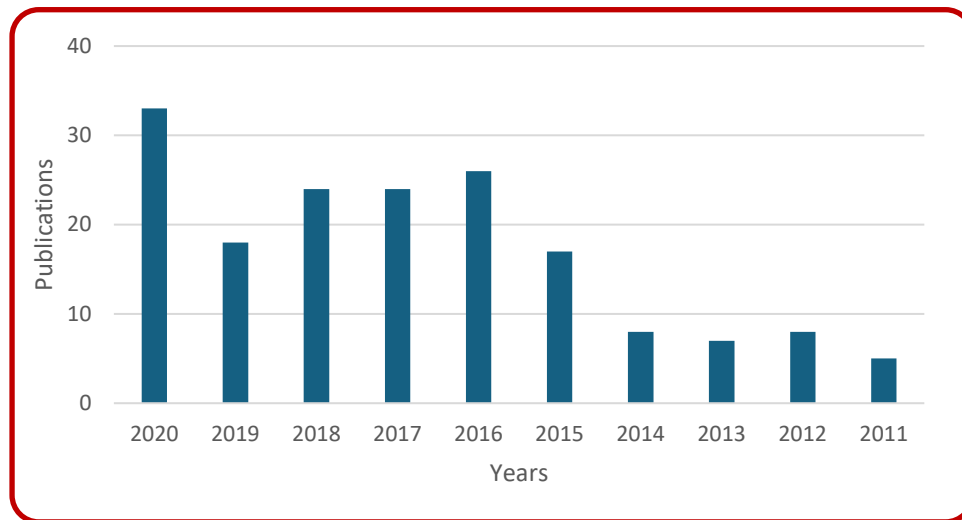


Figure 2. 1: Distribution of annual publications on predicting infectious disease outbreaks research from 2011 to 2020 in Africa.

It can be observed that the prediction of infectious diseases in Africa is a topic that has not been intensively researched. This is demonstrated by 169 papers published in the ten years up to 1 May 2020 and 247 papers published in the previous three decades, compared to a staggering 1880 publications on the same topic globally (Yang, Zhang and Ma, 2020). Africa accounts for less than 14% of the publications in this field. Despite this, there has been a constant growth in publications for the past ten years, except for 2013, 2017 and 2019 where the annual number of publications saw a decrease of 12.5%, 7.7%, and 25%, respectively. On the other hand, there has been a significant increase in publications from 2014 to 2015 to 2016 and from 2019 to 2020. In 2015, there were 17 published documents, representing more than a 100% increase from 2014. In 2020, 33 documents, representing an increase of 83.3% from 2019, were published. The years with the highest and lowest publications are 2020 and 2011, respectively. Overall, there is a slow but notable growth of research on the prediction of infectious diseases in Africa - this is demonstrated by the annual average growth rate of published papers of approximately 23.33%.

There is an indication that the increase in publications concerning the prediction of infectious diseases is associated with outbreaks of infectious diseases. This is illustrated in Table 2.1 below, representing the trends of severe outbreaks or first-time detections of infectious diseases in Africa. For instance, as mentioned earlier, there was a significant increase in 2015, 2016 and 2020. During this period, some parts of Africa experienced the largest outbreaks of infectious diseases. For instance, from 2015 to 2016, Angola and the Democratic Republic of Congo experienced a large outbreak of yellow fever and a large monkeypox epidemic was reported in Nigeria (Fenollar and Mediannikov, 2018).

Moreover, from 2013-2016, the world's largest Ebola outbreak was reported in several African countries, such as Sierra Leone (World Health Organization, 2014). In 2017, the worst plague outbreak of the 21st century was reported in Madagascar (World Health Organisation, 2017). There was also increased interest in researching the prediction of infectious diseases in Africa in 2020 (indicated by many publications that year), at the time Africa and the world at large were experiencing the outbreak of COVID-19. However, few 2020 papers address COVID-19 in Africa. The other factor that boosts the number of publications in the prediction of infectious diseases in Africa is the re-emergence of infectious diseases (Yang, Zhang and Ma, 2020).

Table 2.1: Severe outbreaks / first outbreak detection of infectious diseases in Africa from 2010 to 2020.

Infectious Disease	Period of occurrence	Countries / regions	Impact	Sources
Plague	2017	Madagascar	2 348 cases, 202 deaths	(World Health Organisation, 2017)
Measles	2010-2013	DRC	Largest outbreak: 294 455 cases, 5 045 deaths	(World Health Organisation, 2019)
Yellow Fever Virus	2015-2016	Angola, DRC	Largest outbreak: 7 334 suspected cases, 393 deaths	(Kraemer <i>et al.</i> , 2017)
Ebola	2013–2016	Guinea, Sierra Leone, Liberia	Largest outbreak: 28 646 cases, 11 323 deaths	
Monkeypox	2017	Nigeria	Largest outbreak: 146 suspected cases, 42 confirmed cases, 1 death	(Yinka-Ogunleye <i>et al.</i> , 2018)
Zika Virus	2015-2016	Cabo Verde	First outbreak detected: 7 580 suspected cases	(Lourenço <i>et al.</i> , 2018)

COVID-19	2019 - 27/02/2021 (ongoing)	Africa	3 290 637 confirmed cases, 82 281 deaths	(World Health Organisation, 2021)
----------	-----------------------------------	--------	---	--------------------------------------

Various countries have contributed publications on predicting infectious diseases in Africa. Based on the first author's country affiliation, the list of ten countries with the most publications is illustrated in Figure 2.2 below.

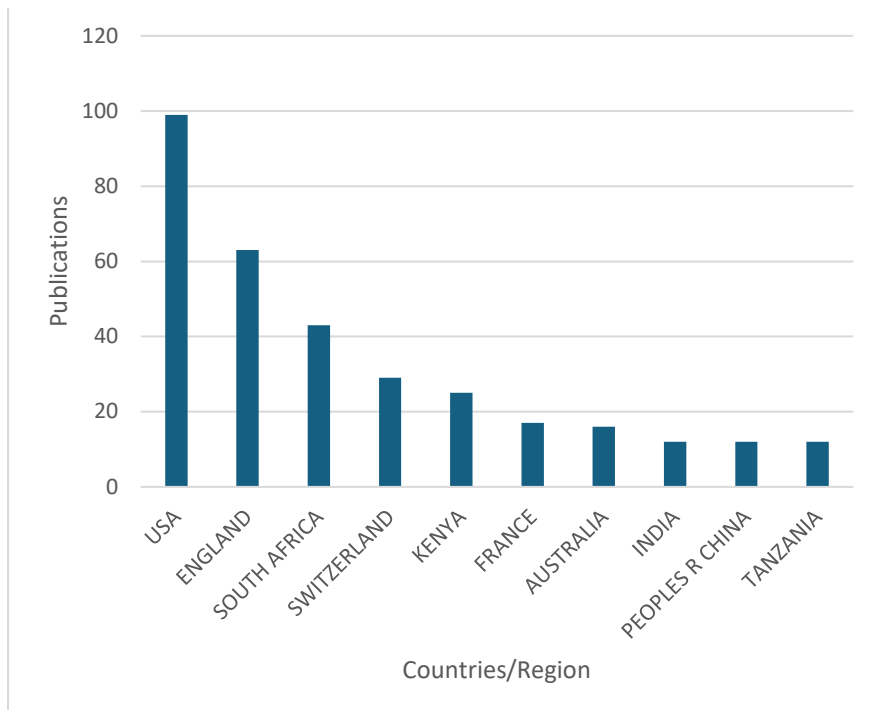


Figure 2.2: Countries in the top ten ranks in publications on predicting infectious disease outbreaks from 2011 to 2020.

The United States of America leads the list, with 40.1% of publications. The second, third and fourth countries on the list are England, South Africa and Switzerland with 25.5%, 17.4%, and 11.7%, respectively. Countries from Europe and Africa dominate the list. South Africa, Kenya and Tanzania are the most productive countries in Africa.

2.2.2 Network Analysis

The bibliometric analysis was conducted using two rigorous software packages: VOSviewer (version 1.6.15) and the bibliometric R package. In VOSviewer network visualizations of the items being analysed are represented using circles and the lines interconnecting the items indicate their relatedness. The bigger the circle, the more important or prominent the item is in the research field (van Eck and Waltman, 2010).

cluster include malaria, mortality, regression, model, disease, schistosomiasis, disease mapping, Sub Sahara Africa (SSA) and variable selection. The next largest is the green cluster, which contains keywords such as climate change, temperature, El Nino, rainfall, rainfall and prediction. The blue cluster is dominated by keywords such as Africa, early warning system, interventions and patterns. The yellow cluster is dominated by keywords such as transmission, outbreak, Ebola, COVID-19, China, model and epidemic. The purple cluster includes keywords such Kenya, West Africa, risk, virus, seasonality and mosquito.

Prominent keywords, represented with bigger circles, reflect research direction and active hot topics in researching infectious disease prediction in Africa. These include Africa, malaria, transmission, climate change, model, COVID-19, epidemic, outbreak, impact and prediction, which appear in different clusters, suggesting that these keywords have gained widespread attention from researchers. Keywords with lower frequencies are represented by small circles, representing areas of research which are less common or are emerging/declining areas. These include disease mapping, regression, variable selection, association and Zika Virus.

Transmission and malaria are the keywords that appear the most. Furthermore, most keywords are meteorological terms, suggesting that the prediction of infectious diseases in Africa is weather based. It also emerges that most of the research in this field has been done in Sub-Saharan Africa and Kenya. There are also keywords such as children, which could be because children are usually most affected by infectious diseases in Africa (World Health Organisation, 2020). For example, the WHO 2020 malaria report indicates that about two-thirds of Africa's malaria-related deaths are among children. The situation is particularly alarming in SSA (World Health Organisation, 2020), which may explain why more research has been done on SSA. Furthermore, climate change is also evident in disease prediction research, alongside associated keywords such as model.

2.2.2.2 Thematic Evolution Analysis

Bibliometric R was used to perform thematic evolution analysis. Figure 2.4 illustrates a thematic map showing the key research themes of the published papers on infectious disease prediction in Africa.

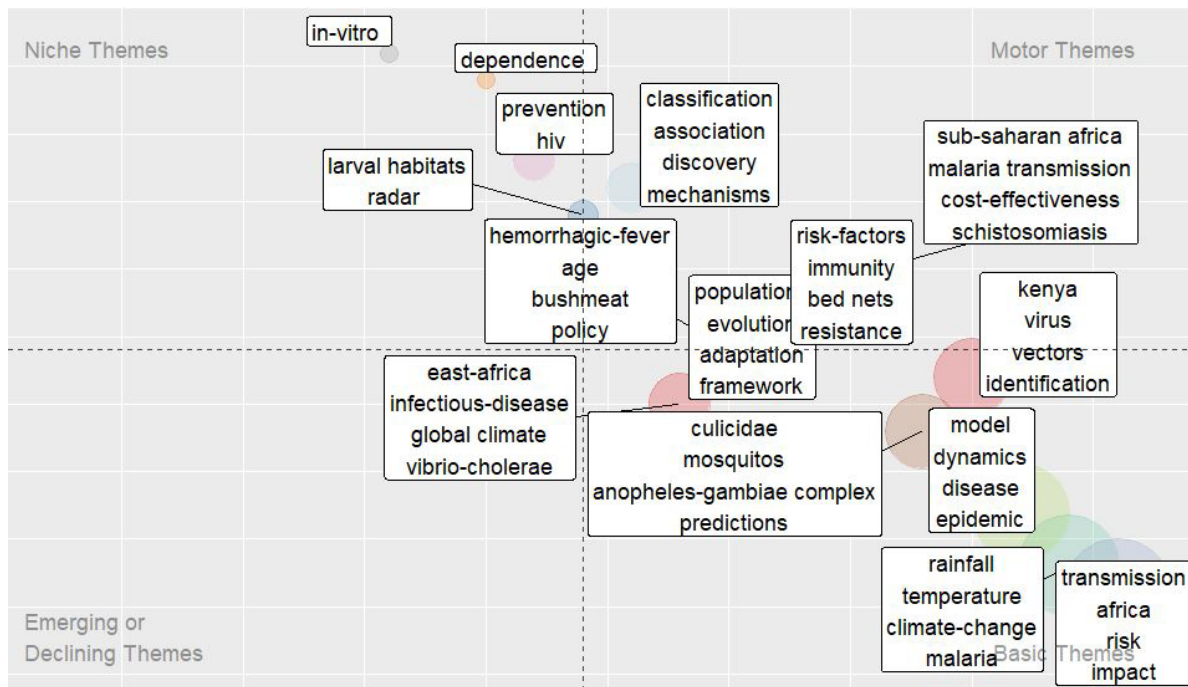


Figure 2.4: Thematic map of emerging themes in predicting infectious diseases research in Africa.

A four-quadrant model is provided, with centrality on the x-axis and density on the y-axis (Cobo *et al.*, 2011). Circle dimensions indicate the number of published documents corresponding to each keyword, i.e., the bigger the circle, the higher the number of publications featuring that keyword in each quadrant (Chen *et al.*, 2019). The motor themes on the upper right quadrant have strong centrality and high density. These are the themes that are considered well-developed and important for the structuring of a research field. These themes can be seen as hot spots in a specific research field. In the research of predicting infectious diseases in Africa, the hot topics include association, classification, discovery, malaria transmission, Sub Saharan Africa, policy, COVID-19 and risk factors. The themes in the lower right quadrant are basic and transversal and considered important but emerging themes that are not well developed. These include keywords like prediction, rainfall, temperature, climate change, model and infectious diseases. It can be observed that most of the basic topics about infectious disease prediction in Africa are not well developed; this is indicated, for example, by most of the keywords with the highest frequency (as shown in Figure 2.4 above) appearing in the lower right quadrant. On the other hand, themes appearing in the lower left quadrant are considered weakly developed themes and have less importance. This indicates that the themes are either emerging or

disappearing topics. This quadrant is empty for the prediction of infectious diseases in Africa. The upper left quadrant contains niche themes. These are the themes with high density but low centrality, and even though they are well developed, they are of marginal importance. These include the following topics: HIV, dependence and larval habitats. Generally, Thematic analysis indicates that research on infectious disease prediction in Africa primarily employs association and classification models, along with climate variability data, to forecast infectious diseases such as schistosomiasis, with a particular emphasis on malaria transmission.

The bibliometric analysis has determined the state-of-the-art of infectious disease prediction research in Africa, highlighting a paradigmatic shift, current challenges and research opportunities. From the bibliometric analysis, the following stood out: (1) severe outbreaks of infectious diseases in Africa during the past decade have been followed by increased scientific publications, (2) malaria prediction is the dominant research theme, (3) research hotspots include malaria, models, classification, associations, COVID-19, and cost-effectiveness, (4) as in other fields, Africa has not kept up with the pace of research in predicting infectious diseases in terms of research output and the use of fourth Industrial revolution (4IR) tools, (5) Africa has focused on weather-based prediction systems considering mainly meteorological factors. African researchers may find it valuable to explore 4IR tools and consider integrating different knowledge systems to predict infectious diseases like malaria.

Even though bibliometric analysis provided the researcher with a valuable birds-eye overview of the literature, there was a need to conduct a more in depth systematic review. Some literature, such as books, are not well tracked by bibliometric analysis (York University Libraries, 2024). Furthermore, since bibliometric analysis primarily examines the bibliographic characteristics of the literature, a systematic review focuses more on content (van Eck and Waltman, 2010; Malapane *et al.*, 2022; Martins and Von Wangenheim, 2023), these two approaches can be considered complementary. For these reasons, it was necessary for the researcher to conduct a systematic review.

2.3 Infectious Disease Prediction

Infectious diseases are illnesses caused by viruses, bacteria and other microorganisms (World Health Organization, 2023). These include but are not limited to malaria, measles, dengue, Scarlet fever, yellow fever, cholera and chicken pox

(Chae, Kwon and Lee, 2018; World Health Organization, 2019). Infectious diseases can be passed to humans from other humans, animals, or via the environment such as water (Berger, 2017). Malaria and other infectious diseases, such as diarrhoea and tuberculosis, are among the top 10 causes of death in developing countries (World Health Organisation, 2020), as illustrated in Figure 2.5.

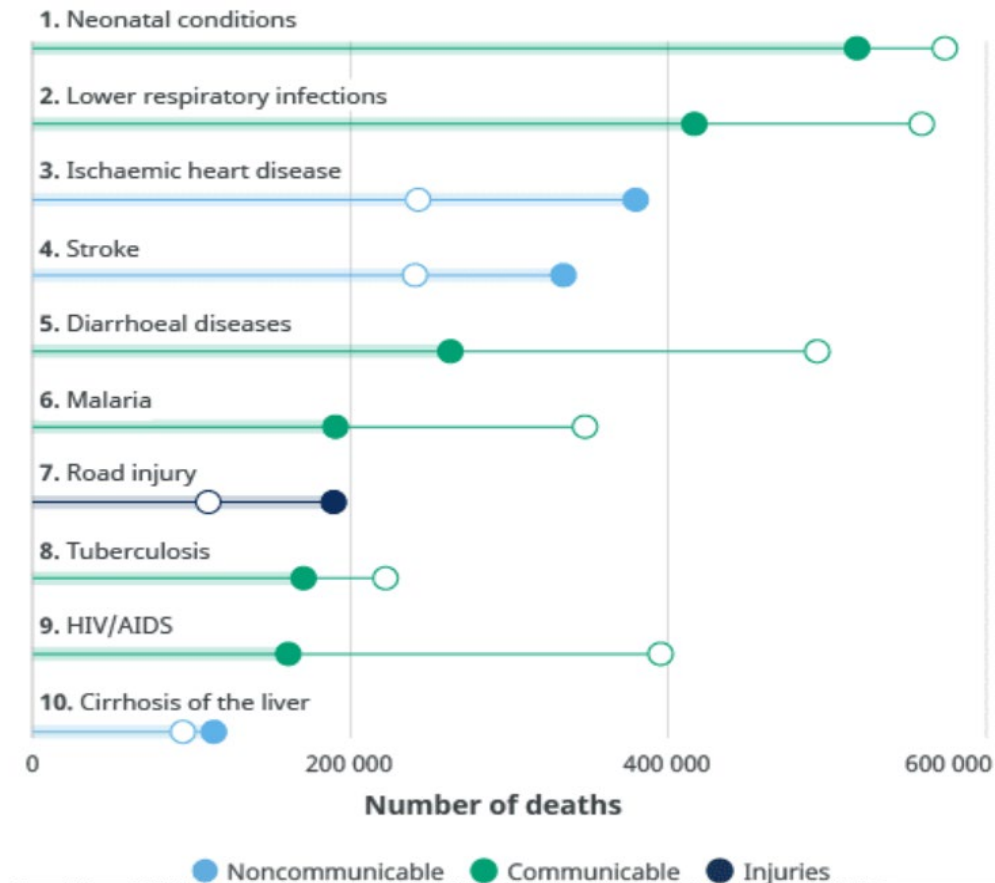


Figure 2.5: Top 10 causes of death in low-income countries from 2000 to 2019 (World Health Organization, 2020b).

Figure 2.5 indicates that infectious diseases cause more deaths in low-income countries than non-communicable diseases and injuries.

For the past decades, methods ranging from traditional to mathematical models and recently ML based approaches, have been used to explore hidden patterns in big data. The aim is to solve, among many others, health issues that arise from climate change (Sharma *et al.*, 2015). For instance, Mohapatra *et al.* (2022) and Masinde (2020) developed ML models to predict malaria outbreaks in India and Africa, respectively; while Ong *et al.* (2023) developed an ML model that considered the climate factors to predict dengue outbreaks in China. Many of these methods assist in diagnosis,

classification and risk forecasting, while few predict future disease outbreaks (Sarkar, 2019). Moreover, most predicting models are based on mathematical models, with a few based on machine learning (Modu et al., 2017). Predicting infectious disease outbreaks is a complex task and incorporating climate data makes it even more complicated because of the dynamic, complex nature of the climate. Thanks to big data and machine learning, it is now possible to map certain climate conditions to specific infectious disease outbreaks.

The section below expands on malaria. Since Africa and in particular SSA is a home for Malaria, it is of interest in this research study to discuss malaria in the context of Africa.

2.3.1 Malaria

Malaria is a vector-borne disease spread by the bite of the infected *Anopheles* mosquito (World Health Organization, 2022). Transmission is more intense in places where the mosquito lifespan is more extended and where it bites humans rather than other animals. The World Health Organisation indicates that the long lifespan and strong human-biting habit of the African vector species are the main reasons why approximately 94% of the world's malaria cases are in Africa.

There is broad evidence linking malaria occurrences with climate variations such as temperature, precipitation, relative humidity, El Niño events and the global vegetation index (Altizer *et al.*, 2013; Amuakwa-Mensah, Marbuah and Mubanga, 2017; Masinde, 2020; Diouf *et al.*, 2022; Woyessa *et al.*, 2023). The correct combination of climatic conditions can create a conducive environment for malaria to thrive because climate changes can impact the survival, reproduction and distribution of disease agents and their means of transmission in the environment (Kurane, 2010; Diouf *et al.*, 2022). There are also socio-economic factors that affect the severity of malaria and its transmission, and these include construction activities, education and population (Sharma *et al.*, 2015; Amuakwa-Mensah, Marbuah and Mubanga, 2017; Cella *et al.*, 2019). Malaria is often prevalent within temperature ranges of 20°C to 30°C (Darkoh, Larbi and Lawer, 2017; Thakur and Dharavath, 2019; Masinde, 2020) and in African regions between latitudes 0° to 15° (Masinde, 2020). However, the relationship between precipitation and malaria is less clear, possibly due to climate change and the interplay of multiple climatic factors. While specific rainfall thresholds vary among studies, it is generally

observed that excessive precipitation can flush away mosquito larvae, reducing malaria cases temporarily, while also creating breeding grounds for mosquitoes in stagnant water afterwards (Githeko and Ndegwa, 2001; Kifle *et al.*, 2019; Sadoine *et al.*, 2024). Other climatic factors, such as relative humidity, impact malaria incidences; an increase in relative humidity can increase malaria incidences (Segun *et al.*, 2020). Models incorporating precipitation and other climatic drivers can successfully predict malaria incidences, although finer-scale modelling at a local level is recommended for more accurate predictions (Kifle *et al.*, 2019).

Other potentially relevant factors such as drought indices have not been explored in predicting malaria, regardless of the remarkable success of utilising these factors in the prediction of phenomena such as drought (Kim, Byun and Choi, 2009; Malik and Kumar, 2020b; Malik *et al.*, 2021). Drought indices quantify drought conditions through the integration of, for example, precipitation, rainfall, stream flow and snowpack data into a logical big picture. Many indices have thus far been developed over the years, including the Effective Drought Index (EDI) (Byun and Wilhite, 1999). EDI considers daily water accumulation with a weighting function for time passage, whereas the Available Water Resources Index (AWRI) is used to express the actual amount of available water (Wilhite and Glantz, 1985). When it comes to malaria transmission, therefore, we hypothesise that the latter could play a role. Integrating data from various sources and employing advanced modelling techniques can enhance our understanding of malaria transmission dynamics and support effective prevention and control strategies like early warning systems.

2.3.1.1 Devastating Impacts of Malaria

The impacts of malaria are devastating. Malaria claims the lives of many people each year globally and Africa continues to bear the burden of morbidity. Countries with a high burden of malaria generally have weak surveillance systems and cannot assess disease distribution trends, consequently, many lives are lost (Amuakwa-Mensah, Marbuah and Mubanga, 2017; World Health Organization, 2020a). Table 2.2 below displays the statistics of malaria deaths in different regions of the world from 2010 to 2021.

Table 2.2: Estimated number of malaria deaths by WHO region (*World Health Organization, 2022*)

	Number of malaria deaths											
	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021
Africa	653000	617000	584000	570000	549000	542000	542000	555000	541000	544000	599000	593000
Americas	502	464	430	470	348	390	530	665	572	510	418	334
Eastern Mediterranean	8600	7800	8000	7500	7500	8200	9500	10200	11100	11900	12700	13400
Europe	0	0	0	0	0	0	0	0	0	0	0	0
South-East Asia	38000	32000	27000	20000	23000	24000	25000	18000	11000	9000	9000	9000
Western Pacific	3500	3000	3500	4000	3800	2500	3000	3000	3000	2600	3200	2600
World (total)	703602	660264	622930	601970	583648	577090	580030	586865	566672	568010	624318	618334
World (children under 5 years)	450000	406000	377000	348000	334000	311000	290000	278000	272000	381000	481000	495000

From Table 2.2, it can be noted that malaria is a global crisis and Africa continues to bear the burden, with nearly one million deaths. For example, in 2021, Africa accounted for about 95% of global deaths of which 80% were children under 5 years old. Nevertheless, generally, there was a slight decline in the number of deaths in Africa until the period 2019-2021. America is the least affected region, while Europe has been declared malaria-free (World Health Organization, 2022).

2.3.1.2 Malaria infection and proliferation in Sub-Saharan Africa

Malaria remains a significant public health problem in many countries in Africa, particularly in SSA, where health resources are usually limited (World Health Organization, 2022). Over 50% of all malaria global deaths in 2022 occurred in four countries in SSA: Nigeria (26.8%), the Democratic Republic of the Congo (12.3%), Mozambique (4.2%), and Uganda (5.1%) (World Health Organization, 2023). In South Africa, malaria is endemic in the low-altitude areas of the northern and eastern parts of South Africa along the border with Mozambique and Zimbabwe, with the transmission taking place mainly in Limpopo, Mpumalanga and Kwazulu-Natal provinces (Moonasar *et al.*, 2021). Limpopo accounts for more than half of the malaria burden in South Africa, with cases mostly occurring in the Vhembe district (National Department of Health South Africa, 2018).

In South Africa, malaria transmission is seasonal, occurring in wetter and warmer months from September to May, with the peak in January to February inclusive (Dieng *et al.*, 2015; National Department of Health South Africa, 2018). In 2018, Vhembe District had a malaria incidence rate of 3.79 cases per 1000 persons (National Institute for

Communicable Disease, 2019; Patrick *et al.*, 2023). *Anopheles arabiensis* and *Anopheles funestus* are vectors that transmit *Plasmodium falciparum*, accounting for malaria transmission in South Africa. *Anopheles arabiensis* is one of the most effective local malaria vectors (Coetzer and Adeola, 2020).

The strategy plan of the National Department of Health to eliminate malaria in South Africa by 2023 has not been successful (National Institute for Communicable Diseases, 2023). According to the National Institute for Communicable Diseases (NICD), some reported challenges included importing malaria cases, vector insecticide resistance and several health provision challenges such as increased disease burden and leadership crisis (National Institute for Communicable Diseases, 2023). On the other hand, the literature points out the lack of effective surveillance as another factor contributing to the unsuccessful malaria elimination plan (National Department of Health South Africa, 2018). At the same time, transmission is generally low and spatially heterogeneous, calling for local malaria intervention, such as at the district or municipal level (World Health Organization, 2017). This is evidence that malaria remains a concern in South Africa.

2.3.2 Infectious Disease Prediction and Information Communication and Technology -Where is the Link?

The contribution of Information and Communication Technology (ICT) to infectious disease prediction systems is apparent, from data collection and integration to data analysis and dissemination of information on infectious disease outbreaks (Adamu, 2021; Alibaba, 2021; Mughal *et al.*, 2022). For instance, Geographical Information Systems (GIS) and phones can be used for data collection. For data integration and analysis, tools such as data mining and artificial intelligence techniques like Artificial Neural Networks (ANN) can be utilised (Verma *et al.*, 2020; Phoobane, Masinde and Mabhaudhi, 2022). Mobile phones and web portals are typically used for information dissemination. The ICTs commonly used to predict infectious diseases include AI, Cloud Computing, web portals and mobile phones (Masinde, 2012).

2.3.2.1 Artificial Intelligence

One significant contribution of ICT to infectious disease prediction are AI techniques such as ML and fuzzy cognitive maps (FCMs) (Yadav, Perumal and Srinivas, 2020;

Groumpos, 2021; Nkiruka, Prasad and Clement, 2021). AI solves complex problems by providing computer systems with the capacity to perform tasks that require human intelligence, such as learning and visual perception (Panch, Szolovits and Atun, 2018). Marr(2019), in his book, shared the real-world uses of AI, acknowledging its critical role in different aspects of our lives by addressing humanity's most pressing challenges, such as climate change and some deadly diseases such as cancer. Marr further asserts that the vast amount of data and the massive computing power of electronic machines are the forces that allow AI to thrive. Usually, different IA techniques are applied to model vast amounts of relevant data and AI then learns which conditions are most likely to result in disease outbreaks. For instance, the capabilities of deep learning to add more hidden layers to ANN allow for exploring complex non-linear patterns in the data (Jiang *et al.*, 2017). This feature of deep learning makes computers capable of processing data and training themselves to process and learn from data (Zhang *et al.*, 2019). Consequently, deep learning is generally suitable for predicting infectious disease outbreaks associated with climate factors. On the other hand, FCMs are suitable for representing and manipulating data that has imprecise categorisation and generalisation (Hung, Nguyen and Walker, 2018) .

2.3.2.2 Cloud Computing

Cloud computing plays a vital role in predicting infectious disease outbreaks, from data storage and real-time processing to providing robust machine learning and predictive analytics tools (Al-Turjman, 2023). Cloud computing refers to providing computing services such as storage and applications over the internet (Sunyaev, 2020). Cloud computing can contribute significantly to understanding and tackling infectious diseases when used along with other technologies, such as AI and big data (Al-Turjman, 2023). The capabilities of cloud computing to provide powerful analytics and infrastructure to store massive data such as health information and data concerning the movement of people is a solid platform from which to tackle serious problems of humanity such as climate change and infectious disease (Sharma *et al.*, 2015). Deploying AI on a cloud computing platform can provide more accurate and real-time prediction of infectious diseases (Al-Turjman, 2023). During the COVID-19 pandemic, cloud-based prediction and monitoring solutions were developed to respond to the pandemic. For example, Alibaba developed cloud and AI solutions to predict, with an

accuracy of 98%, the peak, magnitude and duration of the COVID-19 outbreaks in China (Alibaba, 2021).

2.3.2.3 Mobile Phones

Mobile phones are one of the most influential ICT tools in predicting and monitoring infectious disease outbreaks. One pointer to this fact is mobile penetration, which has increased drastically in Africa and worldwide (GeoPoll, 2019; GSMA, 2024). With everyone using their mobile phones online, a rich treasure trove of data is left behind. For instance, epidemiologists can use such data to predict disease outbreaks and capture the spread of disease (GeoPoll, 2019). In 2010, when an earthquake hit Haiti, researchers tracked people's movement in Haiti via their mobile phones, a fact that would become critical in predicting and monitoring the spread of cholera, which broke out immediately after the earthquake (Bate, 2017). By estimating population movement trends, researchers could monitor cholera outbreaks in real-time (Bate, 2017). Moreover, the massive advancement in mobile technology has expanded their computing power, enabling mobile phones to run various AI applications beyond their traditional functions (Marr, 2019). During the COVID-19 pandemic, several mobile apps and mobile-based systems were used to collect COVID-19-related data, detect, monitor and predict its spread, monitor quarantined individuals' movements and disseminate information on the virus (Bassi *et al.*, 2020; Menni *et al.*, 2020). In this context, mobile phones contribute to disease outbreak prediction and monitoring as computing devices, data collection and information dissemination tools.

2.3.2.4 Geographic Information Systems

Geographic information systems (GISs) also play a vital role in predicting infectious disease outbreaks. GISs are computer-based tools that capture, store and analyse geographically referenced information (Chang, 2019). Depending on the purpose, different layers of information can be combined using the GISs to better understand objects concerning their location. This feature of GISs makes them very useful in explaining events and predicting outcomes. In health, GISs have been used to study areas with similar diseases to establish patterns (Salam and Khan, 2016; Lü *et al.*, 2019). GISs obtain data from different independent sources and then overlay the desired information onto maps specific to the disease. Secondly, GISs prepare and analyse data. GISs use aggregation and geomasking to prepare data without revealing the patient's identity, which is very important when dealing with confidential health data.

The prepared data is then analysed using descriptive statistics and tests for clustering to determine patterns in the data. Lastly, the discerned patterns can be used to monitor and predict infectious disease outbreaks, such as where the infectious disease will spread and how much the population will be impacted (Immergluck *et al.*, 2019).

2.4 AI Techniques in Predicting Infectious Diseases Associated with Climate Change

AI is a broad field with roots in philosophy, mathematics and computer science that aims to mimic human cognitive functions (Panch, Szolovits and Atun, 2018). Different AI techniques have different strengths and limitations. As such, each technique is suitable for specific applications. Understanding each AI technique is critical to correctly applying the correct technique for optimal results (Krittawong *et al.*, 2017). Table 2.3 below summarises common AI techniques that are used in disease prediction.

Table 2.3: AI algorithms in infectious disease prediction.

ML algorithm	Description	Examples of application in disease prediction	Advantages	Disadvantages
Intelligent Agents (IAs)	IAs are flexible, autonomous entities capable of making wise decisions based on their experience and built-in intelligence on what action to take and what interactions need to be done with other agents to achieve goals (Garro <i>et al.</i> , 2018).	The multi-agent system was used for early prediction of urinary bladder inflammation disease (Atteya, Dahal and Hossain, 2010).	IAs are suitable for autonomous complex and distributed systems (Garro <i>et al.</i> , 2018).	IAs are complex (Garro <i>et al.</i> , 2018).
Fuzzy cognitive maps (FCMs)	FCMs are knowledge-based recurrent neural networks representing and analysing complex systems using fuzzy logic and cognitive mapping (Papageorgiou <i>et al.</i> , 2009).	FCMs were used for COVID-19 diagnosis (Groumpos, 2021).	FCMs can model and use knowledge acquired from incomplete, imprecise, uncertain or vague data. This allows the processing of human-originated knowledge (Hung, Nguyen and Walker, 2018).	FCMs have no learning capacity (Hung, Nguyen and Walker, 2018).
Artificial Neural Networks (ANNs)	ANNs are highly interconnected computing systems that mimic human cognitive functions (Jiang <i>et al.</i> , 2017). ANNs perform seven categories of tasks: pattern classification, clustering, function approximation, prediction, optimisation, retrieval by content and process control (Naqvi <i>et al.</i> , 2018).	Malaria outbreak prediction (Sharma <i>et al.</i> , 2015; Thakur and Dharavath, 2019; Yadav, Perumal and Srinivas, 2020).	ANNs are powerful due to the black box and versatile learner concepts (Krittana Wong <i>et al.</i> , 2017). ANNs capture linear and non-linear data and is more accurate (Thakur and Dharavath, 2019; Montesinos López, Montesinos López and Crossa, 2022).	ANNs are unable to handle selectivity variance. Does not work effectively on a small number of datasets (Sharma <i>et al.</i> , 2015; Zahra <i>et al.</i> , 2023). The black box concepts do not work when trying to understand the process.
Naïve Bayes	An older classification algorithm that finds the probability of an event occurring given the probability of another event that has already occurred (Vijayashree and Iyengar, 2016; Sailasya and Kumari, 2021). Suitable for text classifying problems (Krittana Wong <i>et al.</i> , 2017).	Used to predict whether a patient had Dengue or not (Arafiyah and Hermin, 2018), but mostly used in heart disease prediction (Gupta <i>et al.</i> , 2020).	Simplicity and stability (Patel <i>et al.</i> , 2017). Performs well on a small dataset for training (Vijayashree and Iyengar, 2016).	Less accurate than ANN (Krittana Wong <i>et al.</i> , 2017).



Boosting Algorithm	Used to create both regression and classification models. Builds strong models by adding several weak models sequentially (González <i>et al.</i> , 2020). The popular boosting algorithms are AdaBoost and Gradient Boost (Onyijen, Ogieriakhi and Olaitan, 2023).	Used to predict malaria outbreaks (Onyijen, Ogieriakhi and Olaitan, 2023).	Not sensitive to noise in the training dataset and can learn complex patterns in the dataset. Additionally, they are fast to train (Onyijen, Ogieriakhi and Olaitan, 2023).	Sensitive to hyperparameters (Onyijen, Ogieriakhi and Olaitan, 2023).
Bagging Algorithms	Builds several models independently on different datasets and aggregates their predictions through a voting process. Models are training in parallel. Bagging algorithms are used for classification and regression problems (González <i>et al.</i> , 2020).	Used to predict malaria in Indonesia (Khoirunnisa and Ramadhan, 2023).	Easy implementation (Onyijen, Ogieriakhi and Olaitan, 2023).	Computationally costly (Khoirunnisa and Ramadhan, 2023).
Support Vector Machine (SVM)	SVM is used for both prediction and classification of objects into two or more categories (Jiang <i>et al.</i> , 2017; Pisser and Schnyer, 2020).	Malaria and dengue outbreak prediction (Sharma <i>et al.</i> , 2015; Guo <i>et al.</i> , 2017; Modu <i>et al.</i> , 2017). (Singh <i>et al.</i> , (2020) used SVM to forecast COVID-19 outbreaks including, globally, the number of confirmed cases, deceased and recoveries.	Performs well for text classification and with non-linear data (Krittanawong <i>et al.</i> , 2017).	SVM is mathematically complex and computationally demanding (Jiang <i>et al.</i> , 2017; Singh <i>et al.</i> , 2020). Prone to overfitting. It also needs a large amount of data samples with few missing data points (Ahmed <i>et al.</i> , 2017).
Reinforcement Learning (RL)	RL is a learning type used in machine learning that is based on trial and error. RL improves or increases models' efficiency (Krittanawong <i>et al.</i> , 2017).	Used along with other techniques.	RL promises to automate a wide range of decision-making and control tasks in safety-critical domains as they enable complex, adaptive behaviour to be learned automatically (Kahn <i>et al.</i> , 2017).	Many RL algorithms are prone to catastrophic failures in safety critical domains (Kahn <i>et al.</i> , 2017; Adamu, 2021).
Decision Trees (DTs)	DTs use a sequence of decision rules, successively dividing data into smaller sets with every successive division resulting in sets containing elements that are more like each other (Vijayashree and Iyengar, 2016; Montesinos López, Montesinos López and Crossa, 2022).	DTs were used to predict malaria and dengue (Tanner <i>et al.</i> , 2008; Phoobane, Masinde and Botai, 2022).	DTs are more powerful for classification problems. DTs are also popular for their simplicity in decision making (Sharma <i>et al.</i> , 2015). Unlikely to encounter overfitting (Krittanawong <i>et al.</i> , 2017). Efficiently handles missing values, which are commonly encountered in clinical studies (Kothari and Dong, 2001).	DTs have severe limitations in their ability to model complex situations, especially when outcomes or events occur over time (Alagoz, 2011) DTs require large data sets for training (Sharma <i>et al.</i> , 2015).



Deep Learning (DL)	DL is the new era of machine learning. The capabilities of DL to add more hidden layers to ANN allows for exploring complex non-linear patterns in data (Montesinos López, Montesinos López and Crossa, 2022).	Used to predict chicken pox, malaria and scarlet fever outbreaks (Chae, Kwon and Lee, 2018; Masinde, 2020).	Multiple hidden layers in DL make computers capable of processing and learning from data and training themselves to process and learn from data. Consequently, this makes DL suitable for image recognition and for the prediction of certain disease outbreaks that are associated with climate factors (Zhang <i>et al.</i> , 2019).	Due to multiple hidden layers and many parameters, DL is prone to overfitting. Does not perform well on a small training dataset (Krittanawong <i>et al.</i> , 2017).
Rough Set Theory (RST)	RST refers to a new mathematical tool (after probability theory, fuzzy set theory) used for the analysis and knowledge discovery from indiscernible incomplete information (Orlowska, 2013).	RST was used to predict hand, foot and mouth disease outbreaks but did not perform well (Liao <i>et al.</i> , 2017).	The strength of RST comes from its ability to handle imprecise and incomplete information like that which humans must deal with. RST is also simple and easy to use (Han, Zhang, and Wen, 1999).	Classical RST can only process discrete data (Han, Zhang and Wen, 1999).
Markov Decision Process (MDP)	MDP is the reinforcement learning technique used in machine learning for computers to learn from data on how to do tasks in a human-like manner. The learning agent learns about the environment and then makes the best possible decision based on the current state of the agent (Alagoz, 2011).	Used to predict High Pathogenicity Avian Influenza (H5N1) outbreaks, disease transmission and detection rates, as well as the locations of undetected infections in the poultry industry (Jewell <i>et al.</i> , 2009).	The power of MDP lies in studying sequential decisions made in uncertain environments (Steimle and Denton, 2017).	MDP has extensive data requirements (Alagoz <i>et al.</i> , 2010).
Self-Organising Maps (SOPs)	SOPs are types of ANN trained using unsupervised methods. They are suited for finding clusters within data (Mzelikahle and Mapuma, 2018).	SOPs were used in an ensemble intelligent system to predict venous thromboembolism (Sabra <i>et al.</i> , 2020).		Have limited application because of topological defect that makes SOPs settle for a local optimum (Mzelikahle and Mapuma, 2018).
Supervised Learning (SL)	SL uses a dataset labelled by the human, to predict the desired outcome (Krittanawong <i>et al.</i> , 2017). Classification and Regression algorithms fall under SL (López, López and Crossa, 2022).	SL is used along with other techniques.	SL is useful for classification and regression problems.	Since humans label the dataset, SL takes much time and requires a big dataset for learning. Small training datasets can result in inaccurate predictions (Krittanawong <i>et al.</i> , 2017).
Unsupervised Learning (UL)	UL uses unlabelled data to predict the known outcome. Clustering and association algorithms fall under UL (Krittanawong <i>et al.</i> , 2017).	UL is used along with other techniques.	UL is suitable for complex tasks such as robots, speech and pattern recognition applications (Krittanawong <i>et al.</i> , 2017).	UL may experience problems identifying the initial cluster pattern, potentially leading to bias (Krittanawong <i>et al.</i> , 2017).

For ML algorithms, accuracy, precision and reliability are key factors for rating performance (Sarkar, 2019). Performance denotes the proficiency of ML models in precisely mapping input space (infectious disease predictors) to output space, infectious diseases occurrence (Engelbrecht, 2007). Engelbrecht claims that performance quantifies the accuracy of predictions or classifications made by a model on unseen data. Additionally, for the optimal choice of ML technique, factors such as data characteristics, number of features, training time and learning curve are important considerations (Engelbrecht, 2007; Krittanawong *et al.*, 2017).

When evaluating the performance of ML algorithms, it is essential to distinguish between regression and classification ML, as they have different objectives (Chicco, Warrens and Jurman, 2021). Regression ML aims to predict a continuous target variable (Lio and Liu, 2020). Common performance metrics used in regression modelling include Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE) and R-squared. These metrics help assess the accuracy and reliability of regression ML algorithms (Parbat and Chakraborty, 2020).

The MAE calculates the average absolute difference between the actual and predicted values. It quantifies the average magnitude of residuals, the difference between actual and predicted values, in the dataset (Chicco, Warrens and Jurman, 2021). Equation 2.1 is used to compute the MAE.

$$MAE = \frac{1}{n} \times \sum_{i=1}^n |y_{actual_i} - y_{predicted_i}| \quad \text{Equation 2.1}$$

Whereby, y refers to the dependent variable or the target variable in a dataset, which is being predicted. The MSE calculates the average of the squared difference between the actual (y_{actual_i}) and predicted ($y_{predicted_i}$) values across n observations. It measures the variance of the residuals (Chicco, Warrens and Jurman, 2021). Chicco, Warrens and Jurman use the square root of the MSE (RMSE) to obtain the standard deviation of residuals. Equations 2.2 and 2.3 below compute MSE and RMSE, respectively.

$$MSE = \frac{1}{n} \times \sum_{i=1}^n (y_{actual_i} - y_{predicted_i})^2 \quad \text{Equation 2.2}$$

$$RMSE = \sqrt{\frac{1}{n} \times \sum_{i=1}^n (y_{actual_i} - y_{predicted_i})^2} \quad \text{Equation 2.3}$$

The coefficient of determination, often denoted as R-squared (R^2), signifies the fraction of the variance observed in the dependent variable that can be accounted for by the linear regression model (Engelbrecht, 2007; Parbat and Chakraborty, 2020; Chicco, Warrens and Jurman, 2021). This metric is scale-free, meaning that regardless of the magnitude of values – whether small or large – R-squared always remains between 0 and 1 (Theocharides *et al.*, 2018). Equation 2.4 depicts the calculation of R^2 .

$$R^2 = 1 - \frac{\sum (y_{actual_i} - y_{predicted_i})^2}{\sum (y_{actual_i} - \text{meanvalue}_i)^2} \quad \text{Equation 2.4}$$

Whereby meanvalue represents the arithmetic mean (average) of the actual values (y_{actual_i}) across all observations in the dataset.

In classification modelling, the goal is to categorise data points into classes (Alqaissi, Alotaibi and Ramzan, 2022). Performance is often evaluated using a confusion matrix, as seen in Table 2.4, leading to metrics such as: accuracy, which measures the proportion of correctly predicted instances; precision, which assesses the accuracy of positive predictions; and recall, which evaluates the ability to identify all positive instances (Wang *et al.*, 2022). The F1 Score provides a balanced measure of precision and recall, while Area Under the Receiver Operating Characteristic Curve (AUC-ROC) offers insight into a model's ability to distinguish between classes across various thresholds (Alqaissi, Alotaibi and Ramzan, 2022). These metrics provide a comprehensive view of an algorithm's classification performance and its likely ability to generalize to new data (Wang *et al.*, 2022). Equations 2.5, 2.6, 2.7 and 2.8, below, indicate how these metrics are calculated.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FN + FP} \quad \text{Equation 2.5}$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad \text{Equation 2.6}$$

$$\text{Precision} = \frac{TP}{TP + FP} \quad \text{Equation 2.7}$$

$$\text{F1 score} = \frac{2 \times (\text{precision} \times \text{recall})}{\text{precision} + \text{recall}} \quad \text{Equation 2.8}$$

The above metrics are used in conjunction with the confusion matrix in Table 2.4 (Vujovic, 2021).

Table 2. 4: Confusion matrix for binary classification problem (Vujovic, 2021).

Class designation		Actual class	
		True (1)	False (0)
Predicted Class	Positive	TP	FP
	Negative	FN	TN

Confusion matrix has four key components: True Positives (TP), which represent instances correctly predicted as positive; False Positives (FP), indicating negative instances that were predicted as positive; False Negatives (FN), showing negative instances that were predicted as positive; and True Negatives (TN), which are instances correctly predicted as negative (Vujovic, 2021).

The supervised ML models prominent in modelling malaria and other infectious disease outbreaks are discussed in the next sections.

2.4.1 Artificial Neural Networks and Deep Learning

Artificial Neural Networks (ANNs) are large, interconnected networks of computational neurons that present parallel-distributed processing structures (Engelbrecht, 2007). ANNs consist of many neurons organised into layers and connected through weight layers (Pradhan *et al.*, 2020). In its simplest form, ANNs have three layers: input node layer, hidden node layer and output node layer (Pradhan *et al.*, 2020). The typical architecture of an ANN is shown in Figure 2.6.

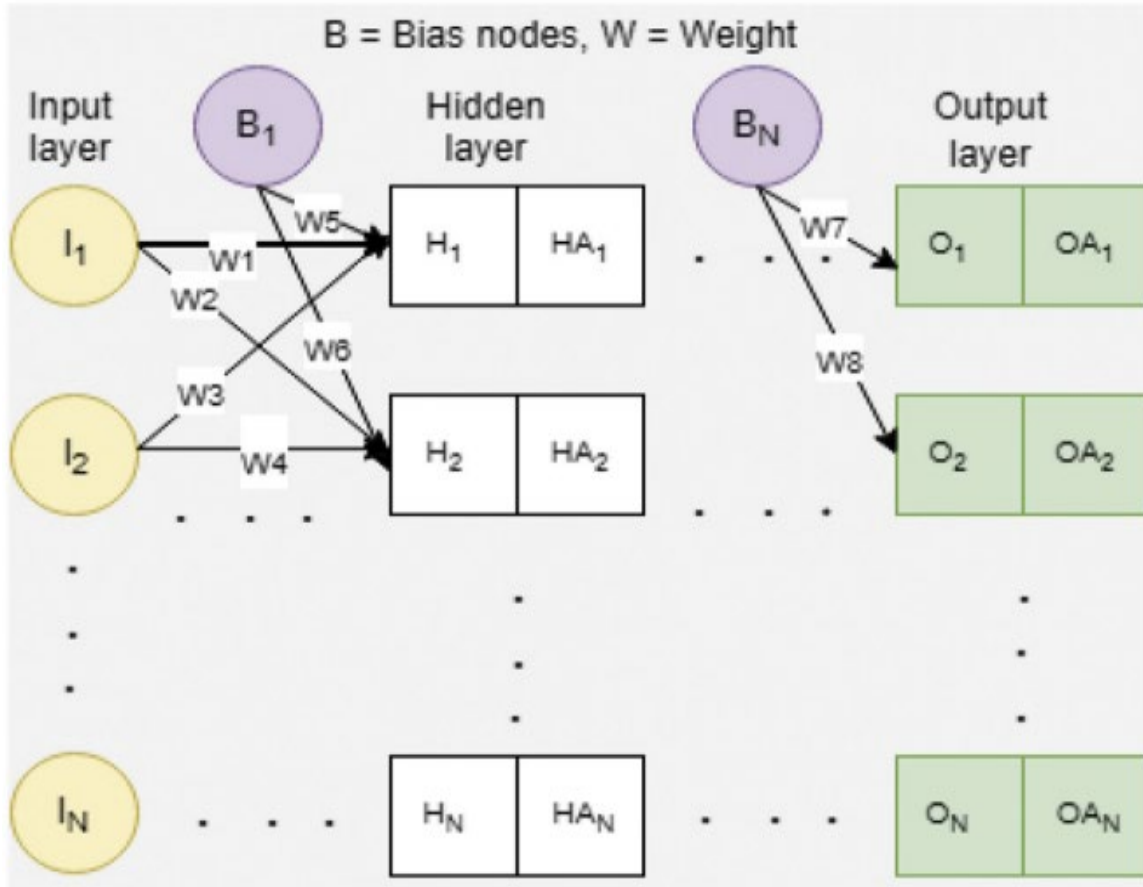


Figure 2.6: ANN basic architecture(Abiiodun et al., 2018).

Figure 2.6 above illustrates a basic architecture of an ANN, consisting of an input layer, a hidden layer, and an output layer. The input layer contains nodes ($I_1, I_2, \dots, I_N$) that hold the input data, where each node represents a feature or variable from the dataset. These input nodes are connected to the hidden layer through weighted connections (W_i). Bias nodes ($B_1, B_2, \dots, B_N$) are present in both the hidden and output layers, adjusting the output along with the weighted sum of inputs to enhance the network's ability to model complex patterns. The hidden layer consists of nodes ($H_1, H_2, \dots, H_N$) that process the weighted inputs from the input layer and apply an activation function to the summed inputs, producing outputs labelled as $HA_1, HA_2, \dots, HA_N$. These outputs then serve as inputs to the output layer, which is made up of nodes ($O_1, O_2, \dots, O_N$) that generate the final predictions. The output nodes receive inputs from the hidden layer, which are weighted by additional weights. The entire network relies on the weights to determine the influence of one node's output on the next node's input, with these weights being adjusted during the training process to minimize prediction errors.

(Montesinos López, Montesinos López and Crossa, 2022). Mathematical representation of this calculation is shown in Equation 2.9 below.

$$v_j = \sum_{j=1}^p w_{ij}x_j \quad \text{Equation 2.9}$$

This equation calculates the weighted sum of the inputs to node j , whereby p represents number of inputs to the node and x_j represents the value of the j -th input feature.

Deep learning (DL) is the new era of machine learning (Gu *et al.*, 2018). The capabilities of deep learning to add more hidden layers to ANN allow for exploring complex non-linear patterns in the data (Jiang *et al.*, 2017; Montesinos López, Montesinos López and Crossa, 2022) explained in Section 2.3.2 (Infectious Disease Prediction and ICT - Where is the link?), this feature makes deep learning suitable for predicting climate sensitive diseases such as malaria.

The remarkable outcomes in perceptual tasks involving abilities that appear instinctive and natural to humans but have long eluded machines are the reasons behind the success of ANNs and DL (Montesinos López, Montesinos López and Crossa, 2022). ANN's strength in solving linear and non-linear multivariate problems has gained a wide range of use in complex applications such as infectious disease prediction. ANNs are popular for their greatest strength in combining various mathematical functions to solve complex problems. As such, ANNs are the frameworks for many different ML algorithms to work together (Sadek *et al.*, 2019; Sharma *et al.*, 2015). Additionally, when used to predict infectious diseases, ANN can take multiple factors such into account such as health data, climate data and vegetable indices (Thakur and Dharavath, 2019).

ANN and DL have been adopted to diagnose and predict diseases, from infectious diseases such as diabetes to chronic diseases such as heart diseases (Sadek *et al.*, 2019; Pradhan *et al.*, 2020). Pradhan *et al.* (2020) used an ANN with Feed-Forward Neural Networks (FFNN) and backpropagation (BPN) as the training algorithm. This model, which was used to predict whether a patient had diabetes, showed good performance with an accuracy of 87.3% and an average error of 0.010. Other projects in which ANNs have been used to predict infectious diseases include the ANN-based prediction of malaria incidences using big data in India (Thakur and Dharavath, 2019).

Moreover, ANN and DL were paramount in detecting, diagnosing and predicting corona virus (Babukarthik *et al.*, 2020; Yang *et al.*, 2021; Absar *et al.*, 2022). Indeed, DL is very good for predictions that are too complex for conventional analysis methods. For even better results, the performance of DL in predictions can be enhanced through parameter tuning (Montesinos López, Montesinos López and Crossa, 2022).

2.4.2 Support Vector Machine

Support Vector Machine (SVM) is an ML method exceedingly popular for its simplicity, flexibility and capability to learn from limited training datasets (Pisner and Schnyer, 2020). It is commonly used to predict infectious disease outbreaks (Singh *et al.*, 2020; Vijander *et al.*, 2020). It is used for both classification and prediction. For classification purposes, SVMs work by categorising data into distinct classes using a hyperplane to define class boundaries, as shown in Figure 2.7.

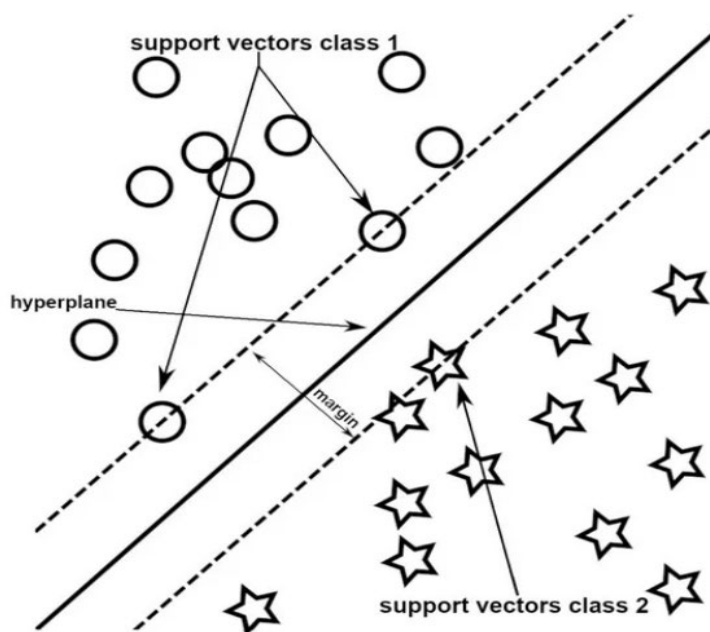


Figure 2.7: A simplified version of SVM (Kufel *et al.*, 2023).

Each data point in the training dataset is assigned to a class. SVMs have two types of lines: boundary lines to create a margin and hyperplane. Support vectors, representing key data points, may lie on or outside boundary lines. On the other hand, a hyperplane is used to create a clear separation of data points into well-defined classes. This hyperplane maximises the distance between itself and the closest data points of each class, thus minimising the classification error (Cao *et al.*, 2003). This way, a clear classification rule is created from the training dataset so that the new data point can correctly be assigned to one of the classes. Support vector regression (SVR) is a

technique based on the same concepts as SVM but used when predicting a continuous value instead of a categorical value (Land Jr *et al.*, 2020).

SVM is popular for its strength in modelling non-linear data (Pisner and Schnyer, 2020). In the studies carried out by Sharma *et al.*, (2015) and by (Modu *et al.*, (2017), SVM was very suitable for malaria outbreak prediction as it was more accurate than other ML techniques. The suitability of using SVM in predicting infectious disease outbreaks was also emphasised by Guo *et al.* (2017) who used SVM to predict dengue outbreaks using climate data. It was found to consistently have the lowest prediction error rates for tracking the dynamics of dengue and forecasting outbreaks in China. Singh *et al.* (2020) also confirmed the high prediction performance of SVM for COVID-19 prediction. However, in a study by Masinde, (2020), SVM did not perform well. For optimal performance, SVM needs a large dataset with few missing data points (Ahmed *et al.*, 2017).

2.4.3 Decision Trees

Decision Trees (DTs) or Classification and Regression Trees (CARTs) are popular for their easy implementation, interpretation and good predictive performance (Modu *et al.*, 2017; Montesinos López, Montesinos López and Crossa, 2022). However, Sharma *et al.* (2015) point out that small data sets used for training can significantly impede CART performance. CARTs build models in the form of an upside-down tree where the root is at the top and the tree branches downwards, with every node representing the decision point except for the leaf.

The CART model uses a recursive binary splitting method to partition the input space (Brownlee, 2016). The process starts with all the data points in one set. Then, the decision tree partitions these data points into two parts using a greedy approach. The greedy approach partitions at the point where the difference between the actual value and the predicted value is minimal. The point at which to partition the datasets is determined by the error measure called residual sum square in a regression problem. Once the partition is done, this process is repeated in each partition until the stopping criteria are met. Stopping criteria are implemented to avoid overfitting (Montesinos López, Montesinos López and Crossa, 2022). When the decision tree method is applied to make a prediction based on an observed data point, the data point is run through the tree from the root by answering the yes/no partitioning questions until it

gets to the last node, known as a leaf. The leaf nodes give a prediction (continuous and/or categorical value) that applies to all instances that reach the leaf (James *et al.*, 2013).

CART is one of the AI tools popular and applicable in health-related studies, including the prediction of diseases (Tayefi *et al.*, 2017). In a study conducted by (Masinde, (2020) to predict outbreaks of malaria, DTs were ranked in the top three out of nine ML models based on computational and classification performance criteria. CART was also used in another study to diagnose dengue, showing a high accuracy of 84.7% (Tanner *et al.*, 2008). CART was also utilised to predict the Zika virus (Bukhari, Webber and Mehbodniya, 2022) and COVID-19 (Groumpos, 2021; Orang, de Lima e Silva and Guimarães, 2022).

2.4.4 Random Forest

The Random Forest (RF) method is a flexible ML algorithm that is easy to use and easily produces competitive predictions of continuous, binary, categorical and count outcomes (Montesinos López, Montesinos López and Crossa, 2022). The capability of RFs to measure the relative importance of each predictor (independent variable) for the prediction has made them popular (Montesinos López, Montesinos López and Crossa, 2022). RF combines the output of multiple decision trees to reach a single result. Some of the attractive features of RF include i) the ability to learn from small training samples yet give competitive results, ii) reduced complexity of computation, iii) ease of use due to less hyperparameter tuning, iv) higher interpretability capabilities and v) low risk of overfitting (Sheykhmousa *et al.*, 2020).

The desirable features of RF have fuelled its adoption in different domains, including the prediction of infectious diseases. In a study by Zhao *et al.* (2020), RF outperformed ANN in predicting dengue, showing low prediction error. RF has also been used to forecast dengue in 790 Brazilian cities with remarkable results (Mussumeci and Codeço Coelho, 2020). In infectious disease predictions, for instance, RF was used to predict African swine flu with a prediction accuracy of 84.46% (Liang *et al.*, 2020), Covid-19 (Yeşilkanat, 2020; Kumar Sharma *et al.*, 2021) and infectious diarrhoea in China (Fang *et al.*, 2020). RF has also played a vital role in malaria prediction, showing high

prediction and classification accuracy (Yadav, Perumal and Srinivas, 2020; Adamu, 2021; Harvey, Valkenburg and Amara, 2021; Mughal *et al.*, 2022).

2.4.5 Fuzzy Cognitive Map

Fuzzy Cognitive Maps (FCMs) are knowledge-based recurrent neural networks representing and analysing complex systems using fuzzy logic and cognitive mapping (Papageorgiou *et al.*, 2009). Kosko (1986) introduced the FCM as a dynamic and semi-quantitative method for structuring expert knowledge. Its origins can be traced back to cognitive mapping, as pioneered by Axelrod (Axelrod, 1976). FCM, a method based on graph knowledge, functions similarly to traditional cognitive maps. Both consist of concepts and their causal relationships. However, FCM distinguishes itself by representing concepts as fuzzy sets and defining relationships through fuzzy connections (Hung, Nguyen and Walker, 2018). Including fuzzy feedback within FCM enables the extraction and modelling of causal knowledge (Gray *et al.*, 2015). By utilizing FCMs, individuals' perceptions can be transformed into conceptual representations and integrated into the system. FCMs facilitate the creation of a broader perspective by incorporating the collective knowledge and insights of the community (Mbele and Masinde, 2020)

An FCM model is a fuzzy directed graph where the nodes (denoted by C) representing concepts or neurons are interconnected by signed, weighted links representing causal effects (Groumpos, 2021). Figure 2.8 shows the structure of an FCM where $C_1....C_6$ denotes concepts while W denotes weight links.

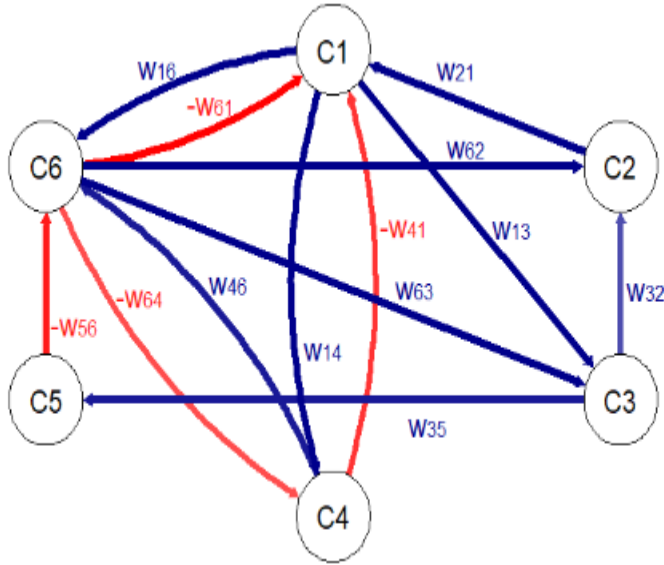


Figure 2.8: FCM structure (Papageorgiou et al., 2018).

The causal effects are crucial as they indicate which concept influences the other and to what degree (positive, negative or zero causality). The causal effect is represented as the numeric weight in the range of $[-1, 1]$ and is commonly depicted in an adjacency matrix (Nápoles *et al.*, 2018; Groumpos, 2021). On the other hand, each concept is characterised by the fuzzy value P (activation value), which represents the probability P of the concept occurrence (Papageorgiou *et al.*, 2009). The bigger the P -value, the greater its impact on the FCM. The activation value, being a probability, is restricted to $P \in [0, 1]$ (Orang *et al.*, 2022). Zero indicates full absence, activation values between 0 and 1 indicates increasing presence, while 1 indicates full presence, as illustrated by equation 2.10 (Nápoles *et al.*, 2018).

$$C(P) = \begin{cases} 1, full_presence \\ \cdot \\ \cdot \\ 0, full_absence \end{cases} \quad \text{increasing_presence} \quad \text{Equation 2.10}$$

The value of each concept, $C(P)$, is influenced by the values of the other connected concepts with their corresponding causal weights and their previous value (Nápoles *et al.*, 2018). An initial activation state of a concept is denoted with C_0 . A new activation vector of concepts is computed at each step t of the simulation by successive multiplications of the state vector by the weight matrix (Papageorgiou *et al.*, 2018), as depicted in equation 2.11.

$$C_i^{(t+1)} = f(C_i^t + \sum_{j=1}^N C_j^t W_{ij}) \quad \text{Equation 2.11}$$

Here C_i^t and C_i^{t+1} are the values of the concept C_i at simulation step t and step $t+1$, respectively. C_j^t is the value of the concept j at simulation step t . W_{ij} is the weight of the interconnection from concept C_i to concept C_j . The inference procedure calculates new concept values during the simulation and continues until the system converges to a steady state (equilibrium), enters a limited cycle or exhibits chaotic behaviour (Nápoles *et al.*, 2018). The function f acts as a threshold, transfer, or activation function to clamp the activation value in the interval $[0,1]$ (Groumpos, 2021; Orang, Silva and Guimarães, 2022).

FCMs have been widely employed in diverse domains to construct monitoring and prediction models (Papageorgiou *et al.*, 2009; Mwagha, 2016; Mbele and Masinde, 2020). Notably, FCMs have demonstrated their efficacy in predicting infectious diseases such as COVID-19 (Groumpos, 2021) and pulmonary infections (Papageorgiou *et al.*, 2009), as well as forecasting drought occurrences (Mwagha and Masinde, 2015; Masinde, Mwagha and Tadesse, 2018). Moreover, Mbele and Masinde (2020) employed FCMs to monitor levels of environmental pollution from mines. Furthermore, there has been a growing trend in applying FCMs for modelling IK in recent years (Sarmiento, 2020; Satama, 2020). Indeed, IK has the potential to tackle challenges such as malaria outbreaks.

2.5 Big Climate Data

2.5.1 Big Data

Before delving into Big Climate Data, it is crucial first to define big data. Big data is an information asset that refers to the data, structured and unstructured, that is so big and complex that it cannot be processed effectively by traditional processing techniques (De Mauro, Greco and Grimaldi, 2016; Manogaran and Lopez, 2017; Somani and Deka, 2017). Big data has a basic characteristic that can be defined by 3Vs: the volume of data, the variety of data types and the velocity at which the data grows (Manogaran and Lopez, 2017). Big data can be sufficiently analysed to get useful insights that can help to understand and solve problems such as climate change, disease diagnosis and disease outbreak prediction (Li and Wang, 2017; Manogaran, Lopez and

Chilamkurti, 2018). The analysis of big data to discover valuable insights is referred to as big data analytics (Bhat, Hegde and Malaganve, 2018; Wang, Kung and Byrd, 2018).

Big data has been experiencing exponential growth recently. The proliferation of big data is due to the massive amount of data in various formats being generated continuously, shared and used in different aspects of life (Manogaran, Lopez and Chilamkurti, 2018). Common data generation sources include pharmaceutical data, electronic medical records, scanned images and financial details (Manogaran, Lopez and Chilamkurti, 2018). One of the elements that fuel the growth of big data is the Internet of Things (IoT) (De Mauro, Greco and Grimaldi, 2016; Ponce and Gutiérrez, 2019). The IoT allows computing devices such as sensors, objects and smart phones to be connected via the internet to collect, process and exchange data automatically (Conti *et al.*, 2018). Another dominant source of big data is social media. As of January 2024, there were 5.35 billion people using the internet, of which 5.04 billion were social media users (Fatehkia, Kashyap and Weber, 2018; Statista, 2024). A mind-boggling figure of 328.77 quintillion (10^{18}) bytes has been estimated as the quantity of data generated daily in 2023 (Duarte, 2024). This break-neck pace is flooding digital devices, especially mobile phones (Goodman, Kamath and Kumar, 2007; Popp and Orloff, 2015; Deloitte, 2019). Figure 2.9 below depicts different sources of big data.

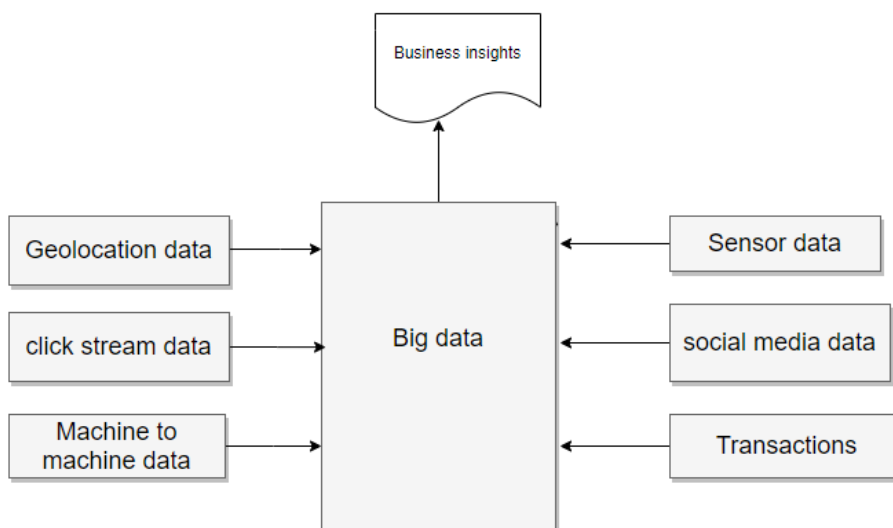


Figure 2.9: Sources of big data (Manogaran, Lopez and Chilamkurti, 2018).

2.5.2 Big Climate Data

In climate science, which is the study of climate (Hu et al. 2018), massive data that is mostly structured has been generated through pervasive computing, AI and IoT weather sensor devices. The latter contributes significantly to generating climate data (Manogaran, Lopez and Chilamkurti, 2018). Consequently, the amount of climate data and the advancement in climate processing technology have grown enormously, giving rise to big climate data. Several organisations, such as World Weather Records (WWR) and the World Meteorological Organization (WMO), have for years been collecting tremendous amounts of daily weather data such as temperature, rainfall and wind speed from thousands of worldwide weather stations (Ceccato et al., 2018).

In the last few decades, there has been remarkable progress in climate science whereby GIS and satellites for earth observation have been developed to monitor the weather, climate, environmental and anthropogenic factors linked to the outbreak of infectious diseases (Ceccato et al. 2018). Ranging from global sources such as Copernicus and Global Earth Observation System of Systems (GEOSS) to country/region-specific weather data providers (such as the South African Weather Service (SAWS)) and freely available weather services (for example, Google Weather App), meteorological centres are continuously generating an unprecedented amount of climate data that cannot be visualised by traditional Geographical Information systems or be stored and processed by traditional data mining algorithms and statistical methods (Manogaran, Lopez and Chilamkurti, 2018). Big climate data can be accessed through the Climate Data Guide (CDG), the Global Climate Observing System (GCOS), EnvironAtlas, the International Research Institute (IRI) data library and the Intergovernmental Panel on Climate Change (IPCC). Even though there is easy access to big climate data, it remains a challenge to process, analyse and visualise heterogeneous, multidimensional large data from different sources (Li and Wang, 2017; Ceccato *et al.*, 2018a). Analysing big climate data is even more complex because of the climate's dynamic and complex nature (Hassani, Huang and Silva, 2019).

2.5.3 Processing Big Climate Data

Performing big climate analytics is challenging since data files are discrete, yet the climate variables must be considered continuous and integrative during the analysis

(Li *et al.*, 2017). Climate data has high dimensionality (space, time plus measured variables), complicating big climate data processing even more (Li *et al.*, 2019). Two operations are performed when processing big climate data: spatiotemporal data mining and spatiotemporal querying. Spatiotemporal data mining focuses on finding interesting patterns or trends hidden in big climate data, such as grouping places with similar climate patterns. When retrieving a part of the climate data, such as average rainfall, a spatiotemporal query can be used (Hu *et al.*, 2017a).

One of the widely used technologies for processing big data is Hadoop. Hadoop is an open source framework specifically designed to deal with Big Data. Hadoop is based on two powerful concepts: the storage components of web crawler Nutch from the first big data processing technology known as ApacheLucene, and the programming model called MapReduce. The power of Hadoop lies in its ability to handle parallel batch processing and its scalability (Preethi and Elavarasi, 2017). The primary components of Hadoop are Hadoop's Distributed File System (HDFS) for distributed storage and MapReduce for data processing (Ciritoglu *et al.*, 2018). Spark is another widely used big data processing tool, which can work as a stand-alone platform or within other applications (Ciritoglu *et al.*, 2018). Spark can be used in Hadoop to replace MapReduce and enhance Hadoop performance due to its superior memory. Other popular tools are Apache Flink, Storm and Samza (Symposium, Engineering and Morais, 2015). Flink can process both streaming and batch data and is often preferred for handling high volume data streams at a low processing latency. Storm, which is compatible with Hadoop, also has tremendous processing power and can process millions of records within a second on one node even though it does not support batch processing. As such, it can be used in distributed real-time processing and in machine learning. Samza is also a tool that processes distributed streamed data and ensures that processed messages are in the same order they were received (Gurcan and Berigel, 2018) Click or tap here to enter text.. Recently, many frameworks based on Hadoop have been proposed for big data management and analytics(Preethi and Elavarasi, 2017). However, some have limitations, as they do not support spatiotemporal data query and analysis (Hu *et al.*, 2017a).

2.5.3.1 *Spatiotemporal Data Mining*

In the last two decades, various analytical tools with high analytical power were developed to process big climate data. However, they only support stand-alone computing systems, impeding the collaboration of scientific research paradigms (Li and Wang, 2017). Data must be managed to allow a distributed parallel process over a cluster of nodes (Hu *et al.*, 2017a). Moreover, analytic tools lack high-performance computing and are not flexible; consequently, their functionalities are static and hard to extend. They also require users to install multiple packages on their workstations and to manually download data. To bridge this gap, big climate data analytics moved to the web to facilitate interactive data analysis (Hu *et al.*, 2018). The web is a typical data dissemination medium and much work has been done to investigate and develop web-based climate analysis systems, out of which various online mapping tools such as NASA WorldWind and Google Earth emerged (Bell *et al.*, 2007; Sheppard and Cizek, 2009) .

Despite the significant contributions brought by these systems, most only act as an online platform to facilitate efficient sharing and dissemination of big climate data; they do not allow data mining. Web-based analysis systems eliminate the constraints suffered by stand-alone systems but they too pose limitations, considering that most support time series in at-most 2-dimensions (2D) (Li and Wang, 2017). Moreover, despite the powerful processing power of some of these web-based analytics tools, they require pre-built operations and their functions and operations are not customisable or scalable (Li *et al.*, 2017).

The emergence of cloud computing and graphical processing unit (GPU) paradigms provides software services and storage for computing big data at a low cost (Zheng *et al.*, 2015). Zheng, Wang and Liu (2015) proposed a framework that processes big data using cloud computing. Climate Analytics as a Service also emerged to provide analytics of big climate data as a web service. This allows high-performance scalable analysis of climate data, which is easily programmable and accessible (Li *et al.*, 2017). There are also powerful tools for spatiotemporal data mining.

2.5.3.2 *Spatiotemporal query*

While powerful technology such as cloud computing plays a vital role in the efficient analysis of big data, efficient data management that queries and retrieves data is crucial in big climate data processing (Appel *et al.*, 2018). Data management systems based on traditional relational databases are limited by low performance when extracting and processing external data (Aydin, Akkineni and Angryk, 2016). Fundamental statistics such as the mean and the location of maxima can be obtained via spatiotemporal aggregation for further analysis. Users frequently have an interest in a specific partition of the data, necessitating spatiotemporal queries (Hu *et al.*, 2017).

In the last few years, multi-dimensional array data models have gained attention when dealing with the 3Vs of big data in disciplines such as climate and health. RasDaMan and SciDB are popular array database query systems that support multi-dimensional arrays with varying structure and size (Appel *et al.*, 2018). However, these techniques were criticised for their lack of high-performance scalability and analysis and the need to convert massive amounts of data from one format to another, which increases the complexity of data management (Aydin, Akkineni and Angryk, 2016). Even though many tools are based on these array-based database systems. Chen, Mao and Liu (2014) argue that a file-based database would be better than an array-based one if incorporated with intelligent caching and indexing. Several studies based on Hadoop MapReduce were carried out to bridge the gap between array-based spatiotemporal data and Hadoop MapReduce (Li *et al.*, 2015). Even though they solved the above-discussed array-based problems, they introduced more overhead problems, such as disk overheads caused by data transformation; in addition, they are not flexible as they require specific configurations for the computing environment (Hu *et al.*, 2017).

Hu *et al.* (2017) proposed an index-based approach to efficiently query and process big array-based climate data without converting data from one format to another or without special configuration of the computing environment, as with array-based data management. The index method is a widely used method to accelerate the query process. One of the limitations of traditional spatial indexes is that they cannot work with data with very high cardinalities. The approach proposed by Hu uses grids to map a file's logical, multi-dimensional array data model to the key values used by

MapReduce. This approach bridges the gap between the array-based data models and block-oriented HDFS storage structures. It enhances MapReduce's processing and querying power by improving locality, whereby computation is carried out on the data where it is stored rather than moving the data to a central computational locale, thus evenly distributing the workload over different nodes in the cluster. Regardless of the significant contribution of this framework to the querying of big climate data, this approach cannot handle data with high spatial resolution.

Li et al. (2017) also proposed another framework that can handle multi-sourced, heterogeneous, massive data sets. This framework also does not convert data. Instead, it stores vast amounts of climate data as distributed files in HDFS in their native nature to provide high scalability and fault tolerance. This approach reported high analytic performance since it uses Hadoop's high performance environment to execute queries in parallel. The proposed framework solves the 3V issues of big climate data as it effectively handles big climate data by providing distributed data storage and advanced analytical operations that were not possible with existing systems. Moreover, this framework can also be used to analyse and query big data in domains other than climate. However, the proposed framework has limitations, as it does not support data sets with different spatial resolutions. More grid transformations and processing models, such as Spark, can be developed and added to the framework to support even more complex data analysis of the unprecedented volume of climate data.

There is also a high demand for tools to facilitate today's collaborative scientific analysis by providing an online platform for organisation, transmission and rendering of big climate data. To address this, UV-CDAT, a Google Earth-based visualisation platform, NASA's Giovanni system and Cesium were proposed, along with other solutions (Li and Wang, 2017; Li *et al.*, 2019). On the other hand, recently, DL models such as Convolutional Neural Networks (CNN) are making their mark in spatiotemporal data mining and querying tasks such as predictive learning, anomaly detection and classification (Wang, Cao and Yu, 2022). The incorporation of ML in spatiotemporal data mining and query enhances the processing of big climate data.

2.6 Indigenous Knowledge in Infectious Disease Outbreak Prediction

2.6.1 Definition of Indigenous Knowledge

Indigenous Knowledge (IK), also known as local, traditional, vernacular or folk knowledge (Jiri *et al.*, 2016; Akanbi and Masinde, 2018) refers to informal, unique, traditional information, skills and knowledge existing within and acquired over some time by a group of people indigenous to a particular geographic area (Grenier, 1998). IK is cumulative as it represents knowledge accumulated over years of experience from careful observations and trial-and-error experiments. IK is dynamic and changes after human innovation and interaction with other external knowledge systems. It is expressed through cultural rituals, beliefs, values, community laws, agricultural practices and the uses of specific plant species (Grenier, 1998; Macherera and Chimbari, 2016). The amount and quality of IK can vary among the people living in the same geographic area, based on factors such as age, gender, role and responsibilities, social and economic status, level of curiosity and daily experiences (Kom *et al.*, 2020). For example, elderly people tend to have more IK than younger generations. The reasons behind this include the fact that young people have less interest in IK and/or are being more exposed to modern technologies (Macherera, Chimbari and Mukaratirwa, 2017; Son, Chi and Kingsbury, 2019; Kom *et al.*, 2020).

The local communities use IK for decision-making and survival in natural and human-built environments (Abera, 2020). Generally, local communities make physical observations and use environmental and astronomical indicators such as the patterns and movement of clouds and objects in the night sky, the flowering or flushing of specific trees and plant species, or the behaviour of birds, insects and animals to understand their environment (Masinde, 2014; Son, Chi and Kingsbury, 2019). In this way, they can, therefore, develop mitigation and adaptation strategies to reduce their vulnerability to environmental variability, such as climate change and its impact (Kijazi *et al.*, 2013; Masinde, 2014; Macherera and Chimbari, 2016; Macherera, Chimbari and Mukaratirwa, 2017; Son, Chi and Kingsbury, 2019). As an example, farmers in the Yao ethnic group in Vietnam use the shape and brightness of the moon to predict the weather for the next cropping season (Son, Chi and Kingsbury, 2019). However, IK has been criticised for its incompleteness, imprecision and inaccuracy (Masinde, Mwagha

and Tadesse, 2018; Kom *et al.*, 2020). Nevertheless, many researchers argue that IK is effective to a certain degree, as can also be said about modern science (Akanbi and Masinde, 2018; Mbewe, Phiri and Siyambango, 2019; Son, Chi and Kingsbury, 2019).

2.6.2 Indigenous Knowledge Indicators for Malaria Prediction

There is substantial literature on the use of IK systems (IKS) to predict weather variations such as drought, rainfall and flood (Nyong, Adesina and Osman Elasha, 2007; Kijazi *et al.*, 2013; Armatas *et al.*, 2016; Jiri *et al.*, 2016; Akanbi and Masinde, 2018; Mbewe, Phiri and Siyambango, 2019; Nyetanyane and Masinde, 2020; Leal Filho *et al.*, 2022; Nyetanyane, 2023), with only a few studies focusing on the prediction of disease outbreaks (Macherera, Chimbari and Mukaratirwa, 2017). Nevertheless, if a community can relate weather variations to the occurrence of infectious diseases, that community can predict infectious disease occurrences that are weather related (Macherera, Chimbari and Mukaratirwa, 2017). Therefore, it is evident that IK can be used to predict weather-related infectious disease outbreaks on the basis that local communities can predict weather conditions using IK.

In Zimbabwe, IK was used to predict malaria outbreaks using rainfall predictions because of the established correlation between rainfall and malaria (Macherera, Chimbari and Mukaratirwa, 2017; Thomson *et al.*, 2018; Cella *et al.*, 2019). Other than the variations in temperature and precipitation, local communities in Zimbabwe also used plant phenology and insects as indicators of malaria occurrence. Swarms of butterflies flying from the west to the east in October meant the upcoming of a good rain season and, subsequently, malaria occurrence. Other indicators for malaria prediction were linked to ancestral spirits and included the perceived noise emanating from the mountains, referred to as the "roaring of mountains", as well as certain behaviours exhibited by ostriches. These indicators were not only used to predict malaria occurrence but also to predict the intensity of malaria occurrences. Even though the indicators were reliable to a certain extent, informal observations of indicators were done individually and were not integrated into the conventional health system, diminishing the importance and effectiveness of IK in the prediction of malaria (Macherera, Chimbari and Mukaratirwa, 2017; Kom *et al.*, 2020). The accuracy of some indicators appears to have been affected by climate change but the fact that several researchers (Mafongoya and Ajayi, 2017; Nyetanyane and Masinde, 2020;

Nyetanyane, 2023) continue with the use of IK indicators for rainfall predictions suggests they are still believed to be accurate, at least to some extent.

An extensive list of over 60 IK indicators for malaria prediction in SSA from the literature has been compiled in Appendix A. In addition to the IK indicators specifically for malaria, IK indicators for rainfall and temperature predictions were also considered due to the link between malaria and these weather parameters, as discussed in previous sections. For example, IK indicators for wildfire outbreaks and other issues whose link to malaria outbreaks has not been established were excluded from malaria IK indicators. The indicators were grouped into insects, birds, plants, animals, astronomical, meteorological and other indicators such as human behaviour and spiritual beliefs.

Under insect IK indicators, termites, ants and worms are the most used insect indicators. When seen in abundance, this indicates that a wetter season can be expected as can malaria. Swallows, sparrows, migratory birds, hornbills and coucal birds seem to be the birds commonly used for predicting rainfall and malaria. Generally, the more birds seen around and the more playful they look, the more rainfall and malaria incidences are expected. While many wild birds are indicators for long rains, many domestic birds are used as indicators for short rains. With animal IK indicators, wild animals are the primary indicators of rainfall. The appearance of wild animals around settlements without attacking and in abundance, except for springhare and ant-bear, indicates a good rain season. On the other hand, certain behaviours of domestic animals also indicate a good rainfall season. For example, frequent sight of cattle running with tails raised up indicated rainy seasons and high malaria incidences. Additionally, trees such as Muchakata, gan'acha, mushuku, mango trees, pears, coffee, avocado plums trees, Loquarts and pigeon pea plants are common plant indicators for rainfall. The time of flowering and the number of flowers and fruits the trees bear seem to be critical indicators for rainfall prediction depending on tree type. For instance, if *Acacia nigrescens* bears many fruits, it signifies no rain, while the same behaviour in trees like shepherd's tree indicates abundant rainfall. Heavy flowering of muchakata, gan'acha, mushuku and mango trees signifies a possibility of a season with no rain. In contrast, the flowering of pears, coffee, avocado and plum trees indicate the onset of rain and a good rainy season. An abundance of rain is an indication of upcoming malaria.

The moon, temperature and wind direction are the common astronomical and meteorological indicators for rainfall. On the other hand, the sun seems to be the least used indicator to predict rainfall. It generally seems that the higher the temperature, the more rainfall could be expected. With regards to wind, the wind direction is the critical indicator of rainfall and malaria, with the interpretation of wind direction differing from one community to another. Indicators classified under 'other indicators' were the least used for malaria and weather prediction.

Most indicators seemed to be similar across regions of SSA, with slight differences in interpretations. It is also interesting to observe similar interpretations of IK indicators within Africa and other regions, such as Asia (Son, Chi and Kingsbury, 2019). IK indicators do appear to be useful when predicting rainfall (Mafongoya and Ajayi, 2017).

2.6.3 Infectious Disease Prediction Using IK vs Modern Science

Several researchers concur that IK and modern science complement each other (Durie, 2004a; Smylie, Kaplan-Myrth and Mcshane, 2009; Son, Chi and Kingsbury, 2019). For example, Smylie, Kaplan-Myrth and Mcshane (2009) reveal that local communities in Canada use both local knowledge and modern health services. The younger generation admitted that they would first consult with their elders and then consult with modern health workers to expand their range of options for action. It would usually emerge that both elders and modern health workers diagnosed the patients similarly. On the other hand, in a study by Macherera and Chimbari (2016), the EWS developed using IK was incorporated into a government health plan to fight malaria in Zimbabwe. The system benefited both local communities and the health workers, increasing resilience towards malaria. Health workers worked hand-in-hand with local communities to document observations, which had initially been done informally and individually without any documentation, hindering the effective use of the indicators (Macherera and Chimbari, 2016). Macherera and Chimbari assert that health workers and local communities learnt a lot from each other through the development of an EWS for malaria. It should be noted that most local communities in Africa consult both IK and modern health practitioners to deal with infectious and other diseases. Even though this may be connected to local communities residing in poor rural areas still suffering disparities in modern healthcare delivery (Assan *et al.*, 2009), this is an indication that IK has always worked for local communities, at least to a certain extent, as emphasised by Masinde (2014).

IK and modern knowledge systems can complementarily fill gaps in each other to solve complex problems. IK and modern science-based disease predictions are not mutually exclusive, even though the two have a significant discordance. Table 2.5 shows the complementary aspects between IK and modern science (Tsuji and Ho, 2002; Masinde, 2014).

Table 2. 5: The complementary aspect of IK and Modern Science.

Complementary aspect	Explanation
Diachronic verses (vs) systematic complementarity	While modern science focuses on collecting dynamic data in a time series over a large area, IK focuses on diachronic data in a term series over a small area. For example, the prediction of malaria in a village. The use of modern science and IK can take advantage of the relative strengths of each piece of knowledge, which, in this case, is the general and micro-level focus of modern science and IK.
Averages vs extremes	Modern science systems are based on quantitative analysis of averages, while IK systems are based on observations of extreme weather events and unusual weather patterns. These are both critical in understanding climate change and building, for example, a system that predicts infectious diseases related to weather variations.
Qualitative vs quantitative information	IK systems use qualitative data which is rapid and inexpensive even though lacks precision. On the other hand, modern science systems use precise quantitative data that lacks explanation. Qualitative and quantitative data are vital ingredients to understanding a complex system such as a disease prediction system.
Hypothesis vs mechanisms	While IK systems use hypotheses to solve daily life problems, modern science has powerful tools to provide explanations for the mechanisms.
Objectivity vs subjectivity	Modern science's consideration of facts is not based on people's feelings, while IK's consideration of facts is subjective. Therefore, a system that incorporates both IK and modern science could be seen as science with heart.

Hence, any system that combines both knowledge systems could be more effective.

2.6.4 Differences and Similarities between IK and Modern Science

As mentioned in Section 2.6.1, IK is acquired through historical and current cautious observations of the environment. It is also acquired informally and orally or through imitation and demonstration from the elders by the younger generation (Dei, Hall and Rosenberg, 2000). As such, IK may be seen as subjective and arbitrary. On the contrary, modern knowledge is generated formally by experts using rigorous scientific methods. Therefore, modern knowledge is often considered to be more objective and thorough, with precise measurement and empirical testing of observable events and

trends confirming credibility and legitimacy. In both IK and modern science systems, information is organised to convert experience and beliefs into knowledge. Durie (2004) emphasizes that IK and modern knowledge are each built on distinctive philosophies, methodologies and criteria. As such, the specific criteria used to evaluate formal science or IK should not be used to evaluate the other. Instead, since IK and modern science complement each other, they can be integrated to create new knowledge that reflects dual persuasions to tackle complex challenges such as climate change. There are many similarities and differences between modern science and IK for disease forecasting (Macherera, Chimbari and Mukaratirwa, 2017; Mafongoya and Ajayi, 2017) depicted in Table 2.6 below.

Table 2.6: The differences and similarities between LK and modern science forecasts on disease outbreaks.

IK	Modern Science
It is geographically oriented; it is tailored for a specific social group for a specific time.	Applicable to relatively large areas and lack specificity. It searches for universal validity.
Uses environmental, astronomical, and spiritual indicators to predict disease outbreaks.	Formal models of measurable meteorological and health data and other big data are used to predict disease outbreaks.
Communication is usually oral and repetitive.	Communication is usually formally documented (written or recorded)
Knowledge is passed informally from elders to the young generation and through interaction with the environment.	Knowledge is usually passed formally through lessons, training or reading, generally with a certificate/qualification awarded after learning.
IK has a social context.	No social context.
Information is usually disseminated using dialect.	Information is usually disseminated using language which is unfriendly to local communities.
Data used for prediction is qualitative, such as IK indicators and their corresponding significance.	Data used for prediction is primarily quantitative, such as meteorological data (amount of rainfall etc).
Methods used for prediction are not documented.	Methods used for prediction are documented.
The effectiveness of the methods used for prediction is subjective.	The effectiveness of the predicting methods is explicitly tested and measured.
Experiences, beliefs and information are converted into knowledge	Experiences and information are converted into knowledge.
The explanation is based on spiritual and social values.	The explanation is based on established theory.
IK is concerned with the immediate and concrete necessities of people's livelihoods.	It is concerned with the general explanation obtained from people's daily livelihoods.
IK advances based on the deductive logic of prior achievements.	It advances based on new experiences, not deductive logic.

2.6.5 Indigenous Knowledge and Modern Science Integration

Integrating IK systems with modern science systems dates back many decades (Scheinman, 2000; Mafongoya and Ajayi, 2017). With climate change challenges becoming more complex each time, efforts to integrate different knowledge systems, such as IK and modern science, to combat climate change remain paramount (Mafongoya and Ajayi, 2017). This way, we can construct decision-making systems based on the best available knowledge. Nevertheless, the discordance between these knowledge systems makes the integration process daunting (Masinde, Mwagha and Tadesse, 2018). Integrating IK into modern science systems requires representing IK in a format that modern science systems can process. One of the AI tools that has demonstrated success in processing IK data is FCM, as discussed in Section 2.4.4 (AI Techniques in Predicting Infectious Disease Associated with Climate Change). Systems that integrate both IK and modern knowledge take advantage of the strengths of these knowledge systems to build even more powerful knowledge systems.

Even though there have been initiatives to integrate IK and Modern Science (Masinde, 2014; Akanbi and Masinde, 2018; Mbele and Masinde, 2020; Nyetanyane and Masinde, 2020), such integration has not been explored yet to predict malaria and other infectious disease outbreaks. Therefore, building an EWS that uses both IKs and modern scientific knowledge to predict infectious diseases has a significant untapped potential to tackle malaria.

2.7 Early Warning Systems

Globally, many scientific research and engineering initiatives are underway to track environmental degradation and man-made and natural disasters. These initiatives, usually called early warning systems, aim to predict potential disasters and lessen their devastating impacts on the environment and human life (Quansah, Engel and Rochon, 2010). According to the United Nations Office for Disaster Risk Reduction (UNISDR) (UNISDR, 2006; Luther *et al.*, 2017), early warning systems consist of four components namely:

(1) gathering risk knowledge: At this stage, heterogeneous data is collected and analysed to understand the vulnerability and risk in a certain community.

(2) monitoring and predicting the situation: This is the core of the early warning systems. Relevant parameters must be monitored and a sound forecasting mechanism should be used to generate timely early warnings.

(3) communicating warning messages: The warning message should promptly reach those at risk. The warning message should be communicated in clear, simple language to ensure the community can easily understand it.

(4) responding to the warning: When the epidemic strikes, the appropriate response will be geographically based and disease specific. A wide range of control interventions, with clear roles and responsibilities for those involved, should be put in place. Normally, the response to the epidemic lies in the hands of the government or non-governmental organizations and in some cases may require international intervention (World Health Organization, 2018). The response includes the possible evacuation of people at risk (Ceccato et al., 2018). UNISDR (2006) asserts that the response plans should incorporate local capabilities and knowledge of the at-risk community.

2.7.1 Early Warning System for Infectious Diseases

Several EWSs have been developed to respond to infectious disease outbreaks such as dengue (Shi *et al.*, 2016; Hussain-Alkhateeb *et al.*, 2018), zika virus (Chien *et al.*, 2018) and malaria (Githeko *et al.*, 2014, 2018). Most EWSs are on dengue and mostly conducted in Asia, while zika and malaria EWS are mostly developed for America and Africa, respectively (Hussain-Alkhateeb *et al.*, 2021). For malaria EWS, most studies used weather parameters (from precipitation to temperature) (Hussain-Alkhateeb *et al.*, 2021). Although several studies have developed EWS for predicting infectious diseases such as malaria, these are based only on modern science approaches (Sharma *et al.*, 2015; Sarkar, 2019) and mostly not in the context of Africa. Even though most predictions have high accuracy, they do not incorporate IK into the prediction. Smylie *et al.*, (2004) argue that early warning systems are more effective if they are relevant and understandable to the communities they are intended to serve.

The utilisation of IKS in EWSs developed in the context of Africa has focused on rainfall or drought prediction for agricultural purposes (Akanbi and Masinde, 2018; Masinde, Mwagha and Tadesse, 2018; Nyetanyane and Masinde, 2020; Nyetanyane, 2023).

The use of IK to develop an EWS for infectious disease outbreaks has not been comprehensively explored. To our best knowledge, there is only one EWS based on IK developed to predict infectious disease (Macherera and Chimbari, 2016). It, therefore, occurs that IK is one of the underutilized resources in modern decision-making systems when it comes to infectious disease outbreak prediction. Additionally, although the above-mentioned EWS showed a certain level of effectiveness, it did not incorporate any other rigorous knowledge such as modern scientific knowledge. Incorporating IK with another knowledge system may result in more robust EWSs.

2.7.2 People-Centred Early Warning Systems

People-centred early warning systems involve the end users in developing and employing the EWS (Rahayu *et al.*, 2020). When EWS incorporates the people the EWS intends to serve, it becomes understandable, relevant and sustainable (Smylie, Kaplan-Myrth and Mcshane, 2009). This also increases the acceptance and use of the system because the people feel a sense of ownership of the EWS. Additionally, the community and experts bring different experiences and expertise to enhance the system (Berry, 1994; Macherera and Chimbari, 2016).

2.8 Conclusion

It is important to summarise the wide collection of research studies, theoretical frameworks and scholarly discussions examined throughout this chapter. The bibliometric analysis discussed at the onset of the chapter permitted a comprehensive overview of the literature pertinent to the study. The bibliometric analysis was appropriately suited for this contemporary era of expansive data generated at unprecedented speed. The analysis highlighted several key themes: the swift proliferation of publications, notably in malaria research and the identification of research hotspots, among others. Additionally, the bibliometric review highlighted a significant research gap concerning predicting infectious diseases in Africa, particularly through the lens of the 4IR technologies.

The systematic literature review laid the technological foundation for developing an EWS for infectious disease outbreaks, underpinned by two robust knowledge systems: modern science (encompassing big climate data and ML) and IK. The discussion extended to the characteristics of infectious diseases, with a particular emphasis on malaria and its profound impact on human health, especially in SSA.

Ultimately, the narrative discussed the description and processing of big climate data. The review discussed the link between technological advancements in disease outbreak prediction and some theoretical constructs that were part of the discourse. A comparative analysis of AI tools instrumental in disease prediction identified and explored the most effective AI tools. While efforts to enhance the accuracy of infectious disease prediction tools in Africa are underway, prevailing models predominantly leverage modern scientific techniques, such as ML, or solely rely on IK. Based on existing research, a comparative analysis of IK and current scientific approaches revealed a possible complementary relationship between these knowledge systems. The systematic review emphasised the critical need for an integrated, people-centred EWS synthesising multiple knowledge systems, including IK and modern science. Such an integrative approach promises a more effective, relevant, understandable and sustainable solution for predicting infectious diseases.

Building upon the comprehensive insights derived from the literature review, the following chapter on methodology focuses on the practical application of these findings. The methodology chapter will elaborate on the system framework and the research design, data collection techniques and analytical methods employed in the study.

3 Chapter Three: Research Design and Methodology

3.1 Introduction

This chapter provides a foundation for the study by outlining the theoretical framework and detailing the methodology. Through the presentation of these two elements, a clear connection between the conceptual underpinnings of the study and the methods employed to investigate the research questions and the hypotheses is established.

Firstly, the Malaria Outbreak Early Warning System (MOEWS) as the theoretical framework is discussed, focusing on its architecture, components, key theories and how these aspects inform the present research. The background of the MOEWS as a theoretical lens, developed from previous research conducted by the United Nations International Strategy on Disaster Reduction (UNISDR), is provided. These theoretical perspectives are essential in situating the study.

Secondly, the methodology section details the research design, data collection methods and data analysis techniques which align with the MOEWS theoretical framework. The methodology is critical for explaining how the study was conducted and ensuring the reliability and validity of its finding.

This chapter is organised as follows: Section 3.2 and its subsequent subsections provide a discussion of the theoretical framework, highlighting its relevance to the study. Section 3.3 and its subsequent subsections detail the methodology, including the justification for the chosen research design and methods. Combining the theoretical framework and methodology in one chapter provides an overview of the research foundation and approach, setting the stage for the subsequent analysis and discussion of findings.

3.2 Malaria Outbreak Early Warning System (MOEWS) as a Framework

This section outlines the system framework that aligns with the overall aim of this research study: developing an infectious diseases early warning system for the timely prediction of malaria. The proposed framework is based on key elements and checklists provided by the UNISDR. The MOEWS framework involves the integration of multiple sub-systems, which are combined to form a three-layered early warning

system. MOEWS is a blueprint for implementing a system integrating indigenous knowledge (IK) and machine learning (ML). It leverages indigenous knowledge indicators (IKIs) with weather data, hydrological drought indices and historical malaria data to predict malaria outbreaks using the Vhembe district in Limpopo province of South Africa as a case study. The rest of this section is organised into the following Sub-sections: 3.2.2 presents a discussion on the fundamental elements of an early warning system, while Sub-section 3.2.3 outlines the architecture of the MOEWS framework. Section 3.2.4 provides detailed insights into the features of MOEWS and finally Sub-section 3.2.5 summarises the MOEWS framework.

3.2.1 General Elements of an Early Warning System

According to the (UNISDR, 2006) and Luther et al. (2017), an EWS should consist of the following key elements to be effective: (1) gathering risk knowledge, (2) monitoring and predicting the situation, (3) communicating warning messages and (4) responding to the warning. These elements were discussed in Chapter 2.

Three of these four components were foundational for developing MOEWS (components 1, 2, and 3). As in Masinde's (2012) research to develop an early warning system for drought, named Indigenous and Scientific Drought Forecasting Techniques Integration (ITIKI), the fourth component was excluded as it goes beyond the scope of this study. The main objective was to create an infectious disease early warning system (EWS) and provide alerts to all stakeholders, such as policymakers and authorities, who will develop response measures such as medical services in case of malaria outbreaks. These three components were renamed as follows for this study to suit the goals of developing an EWS for infectious disease outbreaks: 1) infectious disease outbreak risk knowledge, 2) infectious disease prediction and 3) infectious disease outbreak communication and dissemination. Figure 3.1 below depicts these three components adapted for an infectious disease EWS (UNISDR, 2006; Luther *et al.*, 2017).

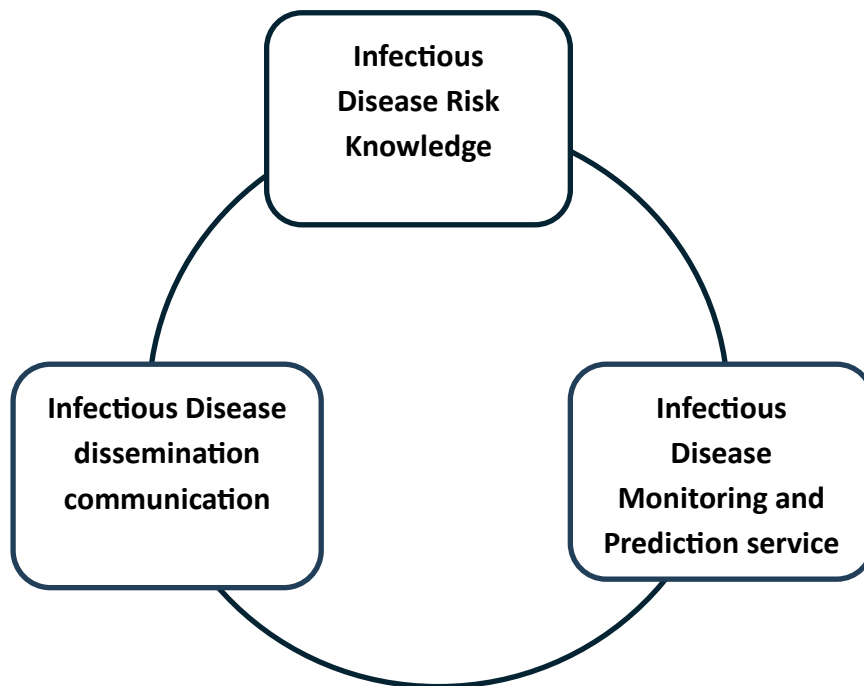


Figure 3.1: The components of an early warning system for infectious diseases.

The first component entails infectious disease data collection and risk assessment. The second component focuses on developing hazard monitoring and early warning services. Lastly, the third component involves infectious disease dissemination and communication. Of all the infectious diseases, malaria was selected because of its devastating impact in Sub-Saharan Africa (SSA). Consequently, the framework was renamed to Malaria Outbreak Early Warning System (MOEWS) and the three components were further adapted for a malaria outbreak EWS. Figure 3.2 below expands on each element.

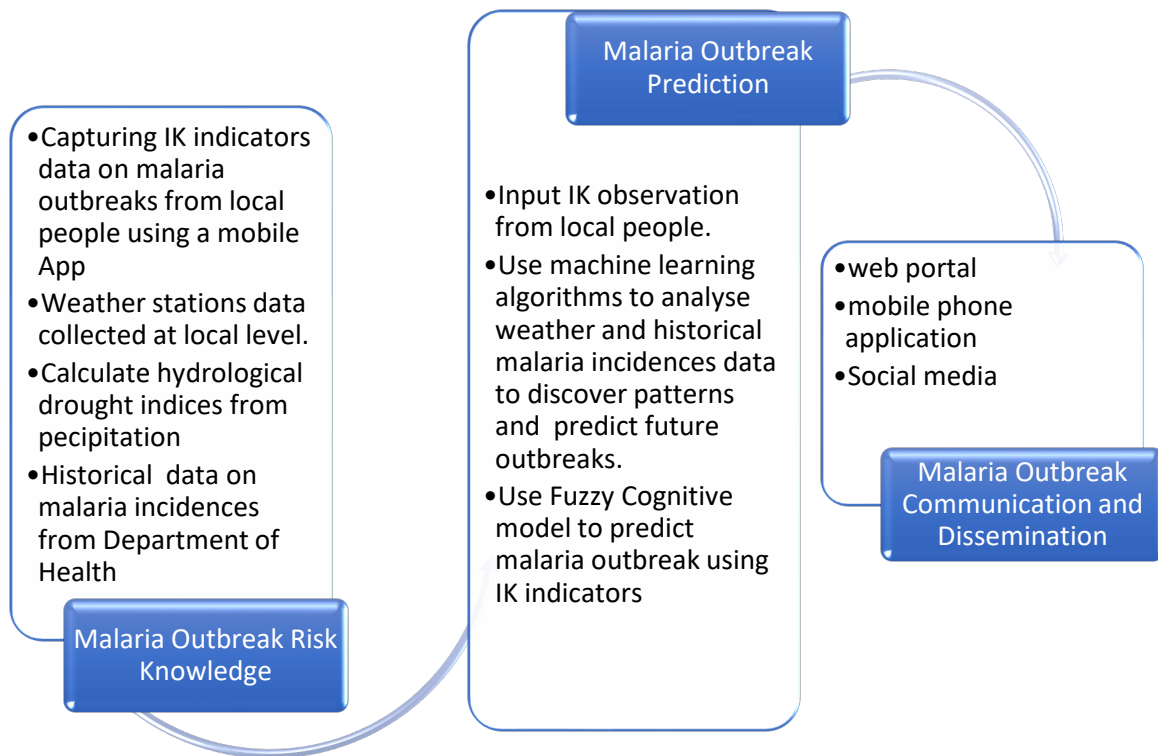


Figure 3.2: Elements of MOEWS.

3.2.2 MOEWS Framework Architecture

In this subsection, details of how each component is supported in the MOEWS framework are described. Figure 3.3 below is a diagrammatic representation of the MOEWS framework.

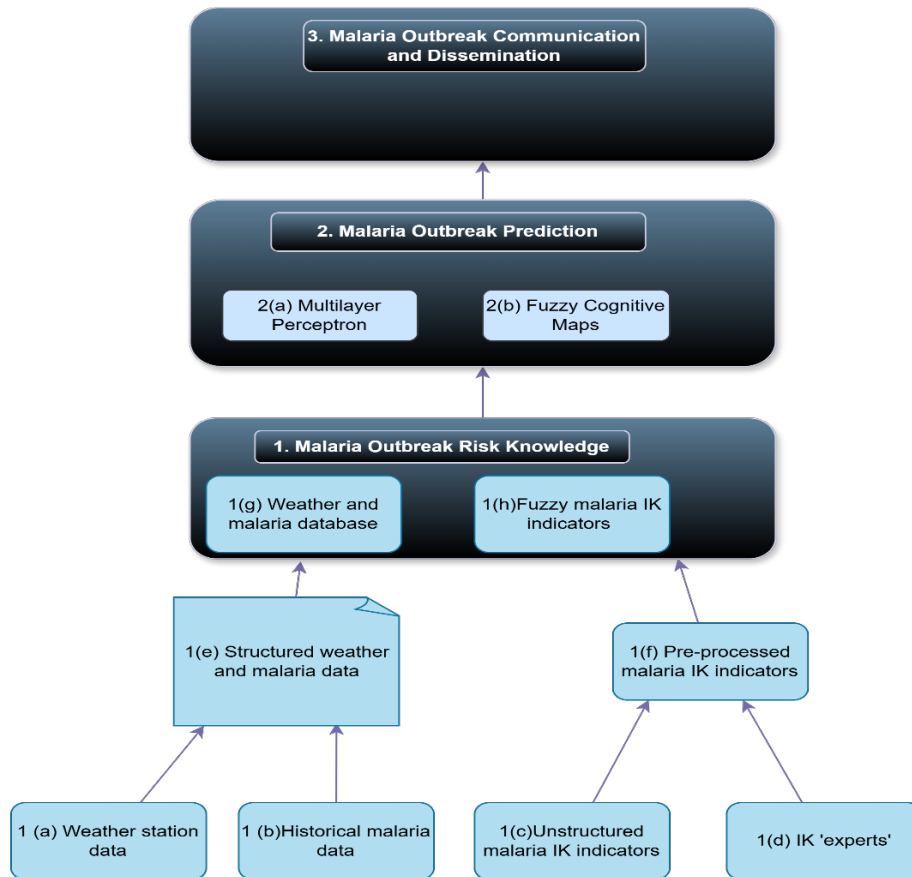


Figure 3.3: Malaria Outbreak Early Warning System Framework.

3.2.2.1 First Component: Malaria Outbreak Knowledge

The first component of the MOEWS framework focuses on malaria outbreak risk knowledge. This component receives data from three domains: conventional weather station data, historical malaria data and indigenous knowledge.

1(a): Weather parameters considered in MOEWS for malaria prediction: precipitation, temperature and relative humidity were collected from the South Africa Weather Station (SAWS). Two hydrological drought indices, available water resources index (AWRI) and effective drought index (EDI) were computed using precipitation data. Weather parameters are pre-processed and held in structured storage 1(e), entered manually into the system, and stored in the structured database in 1(g).

1(b): Structured historical malaria data, sourced from the Department of Health as secondary data, are pre-processed and stored within the same structured repository 1(e) to ensure an integrated system. Data are manually entered into the system and stored in the same structured database 1(g) as the weather data

1(c): In the early stage of this research study, various indigenous knowledge indicators (IKIs) for malaria outbreak prediction were collected from local people in the study area. The IK indicators were verified, pre-processed and stored in 1(f).

1(d): This is the near real-time collection of observed IKIs for malaria outbreak prediction. Local people in the study area, referred to as IK 'experts' or ambassadors, report the IKIs they observe in their environment using a mobile phone application. Only registered users can input the observed IK indicators. The collected IK indicators are pre-processed and stored in 1(f) directly from the phone using the mobile phone's wireless application protocol (WAP) module. The IK experts also report extreme events such as floods.

1(g): Structured database containing weather data, hydrological drought indices and malaria data, and

1(h): The fuzzy presentation of pre-processed malaria indigenous knowledge indicators.

The second component of the framework uses the data collected and stored in the malaria outbreak risk knowledge database to forecast the occurrence of malaria outbreaks.

3.2.2.2 Second Component: Malaria Outbreak Prediction

The second component of MOEWS receives and processes the data from the first component. This component, malaria prediction, plays a role in predicting malaria outbreaks. It consists of two approaches for predicting future malaria outbreaks: Modern Science and IK approaches, which use ML with structured data and fuzzy cognitive maps (FCMs) with IK, respectively.

2(a): Future forecasts of malaria outbreaks are made using an ML model that has been previously trained and tested using historical weather data, hydrological drought indices and malaria data. The predictions are manually stored in the malaria prediction database, as shown in Figure 3.4.

2(b): Using pre-processed observed real-time IKIs and an FCM, which is also a machine learning algorithm, seasonal malaria predictions are made. The predictions are also stored in the malaria prediction database.

The forecasts are sent to the third component of the framework: Malaria Outbreak Communication and Dissemination.

3.2.2.3 Third Component: Malaria Outbreak Communication and Dissemination

The third component of the framework focuses on the communication and dissemination of malaria outbreak alerts as predicted by the malaria prediction component (second component). Malaria outbreak alerts are communicated to the community using a mobile app. Malaria outbreak alerts are also posted on the web portal and social media (for example, Facebook). The web portal and mobile app are also used to output the observed extreme weather events reported by the IK experts in 1(d) in the first component of MOEWS. Below is a system architecture that shows the finer details of the MOEWS.

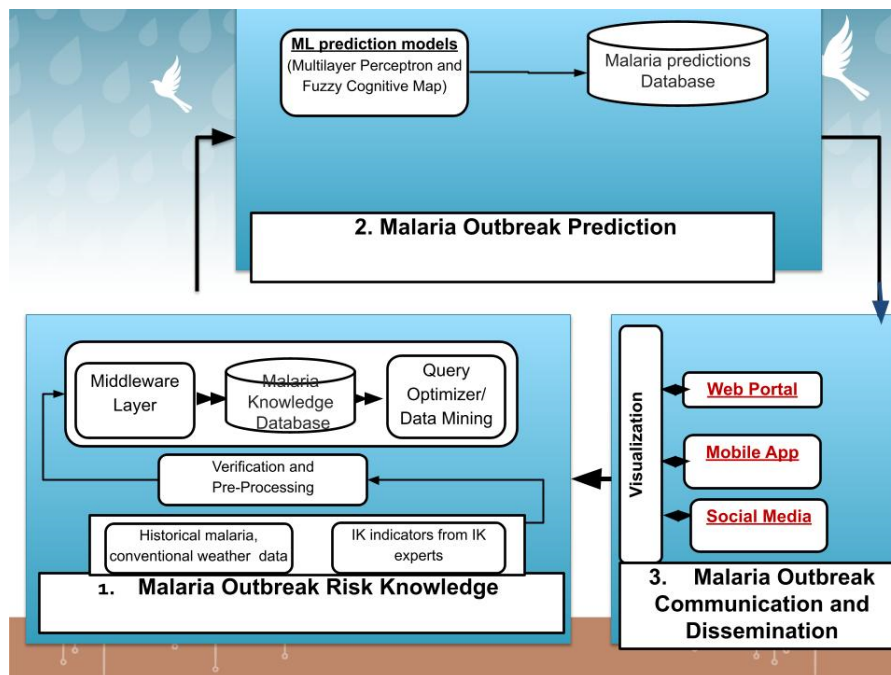


Figure 3. 4: Detailed Malaria Early Warning System Architecture.

3.2.3 Features of Malaria Outbreak Early Warning System

Malaria Outbreak Early Warning System (MOEWS) framework harnesses the power of machine learning in conjunction with IK to develop an intelligent, effective, accessible and relevant framework for combatting infectious diseases, malaria in particular. Through the utilisation of machine learning (ML) models, MOEWS capitalises on the exceptional predictive accuracy demonstrated by ML models in predicting infectious disease outbreaks such as malaria outbreaks (Masinde, 2020; Mussumeci and Codeço

Coelho, 2020; Yadav, Perumal and Srinivas, 2020; Zhao *et al.*, 2020; Adamu, 2021; Hussain-Alkhateeb *et al.*, 2021; Ong *et al.*, 2023). Additionally, incorporating local community knowledge into MOEWS enhances its effectiveness by mitigating several critical challenges that impede the adoption and effectiveness of EWSs. Relying on IK highlights the deeply rooted customs within communities, which could potentially hinder acceptance of external information (De Long and Fahey, 2000). By integrating IK and actively engaging the local community in system development, MOEWS adeptly bridges these gaps. Considering indigenous coping strategies within MOEWS fosters acceptance of MOEWS among local people. Moreover, incorporating IK into the system not only augments community understanding and trust in MOEWS but also promotes its relevance and enhances the ability of local communities to respond effectively to malaria outbreaks. For example, MOEWS specifically incorporates IK indicators used to forecast malaria outbreaks, thereby leveraging the rich predictive capabilities of local knowledge. IKSs have shown considerable reliability in predicting climate-sensitive infectious diseases (Macherera and Chimbari, 2016; Macherera, Chimbari and Mukaratirwa, 2017); hence MOEWS, by considering IK, adds a layer of effectiveness to the framework. Additionally, this approach instils a sense of ownership of the system within the community and facilitates community engagement and empowerment. Notably, EWSs are effective when they are pertinent and comprehensible to the communities they serve (Smylie *et al.*, 2004).

Through the application of MOEWS, observations of IK indicators (IKIs) used to predict malaria outbreaks are reported in real-time via mobile application by registered local individuals (ambassadors or IK 'experts'). The same mobile application is employed to disseminate alerts of malaria outbreaks to communities. The utilisation of mobile applications promotes the accessibility of MOEWS, seamlessly aligning with the growing mobile phone penetration in Sub-Saharan Africa (SSA), as reflected in the literature (GSMA, 2023).

3.2.4 Summary of the MOEWS Framework

Section 3.2 of this chapter presents MOEWS as a framework developed for an early warning system for infectious diseases with a specific focus on predicting malaria outbreaks. The outlined framework, the MOEWS, aligns with the aim of the

overarching research. It leverages two rigorous knowledge systems, indigenous knowledge with machine learning methodologies, to make deductions for inferences regarding malaria prediction. Drawing inspiration from the ITIKI framework, the research study followed a similar methodology as ITIKI to create the EWS framework in the context of malaria outbreak prediction.

MOEWS stands out for its effective, intelligent, relevant and accessible approach to malaria prediction. The system demonstrates intelligent and enhanced predictive accuracy and relevance by incorporating ML techniques, IK approaches and community engagement. A mobile application for real-time reporting of observed IK indicators ensures timely data collection and alert dissemination, promoting effectiveness, accessibility and affordability. MOEWS, as depicted in the framework, illustrating the seamless integration of components for malaria outbreak risk knowledge, prediction and communication.

The MOEWS framework, as articulated above, lays the foundation on which the research design and methodology were structured for this study. The rest of this chapter provides a complete description of the research design and the methodology.

3.3 Research Design and Methodology

This section describes the research design and methodology used in this research study with the previously discussed MOEWS as a theoretical lens. The section is structured as follows: Sub-section 3.3.1 covers the research design, while Sub-sections 3.3.2, 3.3.3, 3.3.4 and 3.3.5 cover the study area, ethical considerations, target population and sampling techniques, respectively. Sub-section 3.3.6 discusses data collection and Sub-section 3.3.7 focuses on data analysis and interpretation. Finally, the chapter ends with a summary in Sub-section 3.3.8. Figure 3.5 summarises the methodology followed in conducting this research study.

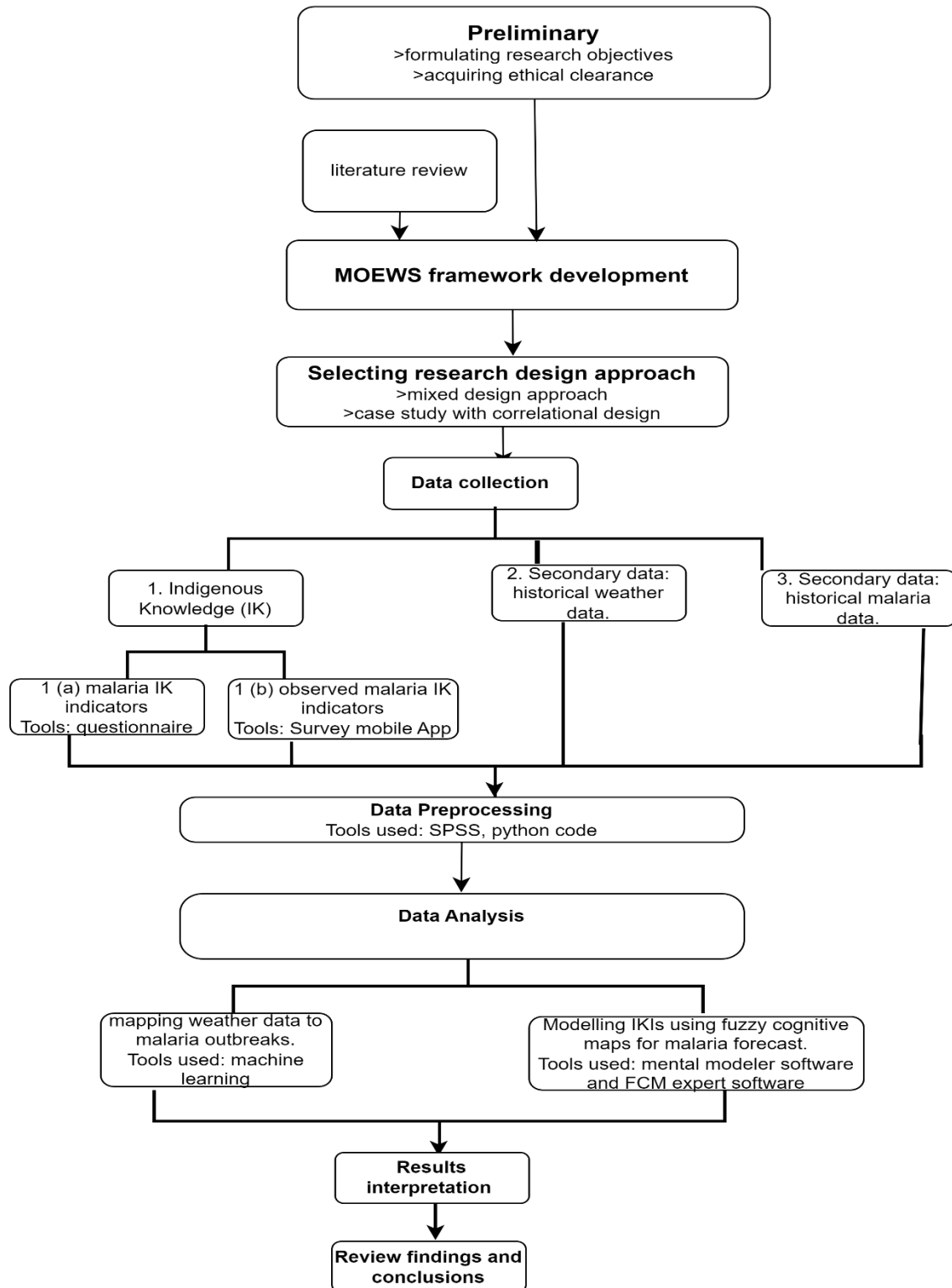


Figure 3.5: Methodology flowchart.

Research can be defined as a scientific and systematic investigation conducted to establish facts and principles in a specific field of knowledge (Cresswell, 2011). Research involves examining and exploring materials and sources to establish facts

and derive novel conclusions (Goddard and Melville, 2004). It follows that to be categorised as research, a study must adopt a systematic and methodical approach, adhering to applicable norms and standards for validity and reliability. Additionally, research should start with clearly articulating the problem that needs to be solved (Hevner, A. and Chatterjee, 2010).

In this thesis, the researcher identified a problem stemming from the challenge that current Early Warning Systems (EWSs) developed for malaria prediction and other infectious diseases are based solely on Western scientific approaches, mostly divorced from the African context. Even though most current EWSs are usually efficient in their prediction accuracy, they can benefit from incorporating indigenous knowledge (IK). Smylie et al. (2004) argue that EWSs are more effective if they are relevant and understandable to the communities they are intended to serve. One way to achieve this in the African context is by incorporating IK into the EWS.

3.3.1 Research Design

Research designs have been growing tremendously due to computer technology advancements (Creswell, 2014; Thomas, 2017). When conducting research in social and behavioural science, the main approaches to research design are qualitative, quantitative and mixed methods research (Tashakkori and Teddlie, 2009). Under the quantitative research approach, several research designs exist that are either experimental or non-experimental (Creswell, 2014). The experimental research design includes true and quasi-experimental research, while the non-experimental design includes a survey and a causal and correlation design (Asenahabi, 2019). Considering these aspects of research and guided by the research objectives and questions, this research study used a mixed methods research design, whereby a case study and correlational design were followed. The following sections expounds upon these research designs and how they were deployed in this research study.

3.3.1.1 *Mixed Methods Research Design*

Mixed methods design integrates qualitative and quantitative design features (Creswell, 2014). Quantitative research, usually considered confirmatory research, is embedded in the rationalism philosophy and uses empirical and systematic methods for collecting and analysing numeric data (Kothari, 2004). On the contrary, qualitative research, usually considered exploratory research, is rooted in a philosophy of

empiricism (Yilmaz, 2013). Qualitative research is a flexible, open and unstructured approach to finding answers to the research questions (Onwuegbuzie, 2014).

The appropriateness of integrating qualitative and quantitative research methodologies in this research study was based on the nature of the data, its collection and analysis. The study utilised both unstructured and structured data and semi-structured and structured data collection methods. Neither qualitative nor quantitative methods alone could satisfactorily address the research questions for this study. Qualitative and quantitative designs each possess inherent limitations, therefore the combination of both serves to mitigate the weaknesses (Östlund *et al.*, 2011; Creswell, 2014).

A qualitative design was used to describe the causal effects between IK indicators and expected malaria incidences. Furthermore, through a developed survey-based mobile application, data on observed indicators of malaria prediction were collected. The data included photographs from the environment, Global Position System (GPS) coordinates of the location of the reporting IK expert and the IK description. At the same time, a quantitative approach was deployed to investigate the suitability of using ML in quantifying the prediction of malaria outbreaks using structured historical weather data, hydrological drought indices and malaria data. The prediction of malaria using ML aimed to determine relationships among the variables used, which called for a quantitative approach (Yilmaz, 2013). Employing a mixed-methods approach assisted the researcher in attaining a deeper understanding of malaria prediction, which could not have been achieved through a singular approach. Ultimately, the mixed methods approach enabled the researcher to adequately respond to all of the research questions in the best way possible.

3.3.1.2 Use of Case Study with Correlational Design

Thomas (2017) defines a case study as an in-depth research methodology that studies an aspect of a case or a small set of cases in a highly detailed manner, resulting in a rich, detailed understanding of the case. Thomas further pinpoints that multiple sources of evidence should be used to gain different perspectives of the case and that there should be characteristics of the case that make it of particular interest or value. To predict malaria outbreaks, this research study adopted a case study methodology to analyse historical weather, malaria data and hydrological drought indices in

Vhembe. Additionally, participants were recruited from Vhembe district to investigate the use of IK for malaria forecasting. A case study was appropriate for providing an in-depth investigation of the intricate complexities of using IK to predict malaria outbreaks. In addition, a correlational research design, which aims at predicting and explaining correlations or relationships between variables, was instrumental in this study in unveiling interacting variables and discerning the nature of their interactions, thereby enabling researchers to predict malaria outbreaks (Seeram, 2019). The correlational design also allowed the researcher to determine the relationship between weather data and malaria incidences and to predict malaria outbreaks.

Vhembe district was chosen as a case study since it has consistently recorded high incidences of Malaria in South Africa (Dieng *et al.*, 2023).

The use of a case study with a correlational design was done to achieve the research objectives outlined in Chapter 1.

3.3.2 Study Area

This research study was conducted in Vhembe, one of the five districts in the Limpopo province of South Africa. Figure 3.6 below depicts Vhembe and its four municipalities: Musina, Makhado, Thulamela and Collins Chabane.



Figure 3.6: Vhembe district map (Vhembe District Municipality, 2024)

Vhembe district shares its north-west borders with Botswana and Zimbabwe and its south-east borders with Mozambique. It covers a total area of 25596km² with a population of 1.39 million. The Vhembe climate is a subtropical climate, with wet and dry seasons (Nenwiini, Kabanda and Bokone-bophirima, 2017). The wet seasons are

warm, while the dry seasons are cold (Musetha, 2016). In wet seasons, the temperature ranges from 16 - 40°C, whereas in dry seasons, temperatures range from 12 - 22°C with an annual mean rainfall of about 500mm. Vhembe receives more rainfall than other districts in Limpopo. Rainfall is experienced mainly between October and March, with the peak in January and February. Floods occur occasionally during the rainy season (Nenwiini, Kabanda and Bokone-bophirima, 2017; Kifle *et al.*, 2019). An abundant amount of rainfall received in Vhembe alongside high temperatures creates conducive conditions for mosquito population growth, ability to bite and longer life spans (Agusto, Gumel and Parham, 2015; Kifle *et al.*, 2019; Masinde, 2020).

3.3.3 Ethical Consideration

Ethics could be defined as how the researcher should behave concerning the rights of participants or those who could be affected by a study (Burgess, 2005). In this study, ethics were considered from when the researcher sought access to the participants to when the researcher collected, analysed and reported the findings. Additionally, prior to the commencement of the research study, the researcher applied for and received approval from the Central University of Technology (CUT) to conduct the research. Subsequently, the ethical clearance letter was issued and can be found in Appendix B. The researcher also took an online module on ethics clearance, Introduction to Research Ethics, at <https://elearning.trree.org/> before conducting surveys (see Appendix C for a certificate of completion).

The fundamental principle guiding the ethical considerations in this study was the acknowledgement and preservation of participants' rights to ownership of the knowledge contributed during the study. The participants were assured that the information they provided remained their intellectual property. On the other hand, a consent form (see Appendix D) was administered to safeguard participant privacy and uphold transparency before their involvement in the research study. This consent form was a crucial document outlining the study's purpose, procedures and potential risks, enabling participants to make informed decisions about their participation. The thoroughness of this consent process was instrumental in respecting the autonomy of participants and aligning with ethical standards as stipulated by Bell, Bryman and Harley (2022).

Furthermore, participants were granted the freedom to pose questions and seek clarification on any aspect of their involvement. This measure ensured that participants provided voluntary and well-informed consent, fostering an environment of trust and transparency throughout the research study. By adhering to these ethical considerations, the study prioritised participant autonomy and privacy, maintaining the integrity of the research process.

3.3.4 Target Population

The target population refers to the entire group the researcher is interested in studying and about which they want to conclude (Flick, 2015). The target population for this research consisted of individuals possessing IK to forecast malaria occurrences in Vhembe districts. While the data about malaria incidences and climate variables were extracted from secondary sources, IK data indicators on the prediction of malaria outbreaks were collected from the participants in the Vhembe district using questionnaires.

3.3.5 Sampling Techniques

Sharma (2017:749) defines sampling as “... *a technique (procedure or device) employed by a researcher to systematically select a relatively smaller number of representative items or individuals (a subset) from a pre-defined population to serve as subjects (data source) for observation or experimentation as per objectives of his or her study*”. A multi-stage sampling technique was employed (Sharma, 2017). Purposive sampling was used in the first stage to select a specific group of respondents with IK (Stratton, 2021). The researcher used a snowball sampling technique in the second stage based on referrals from existing respondents (Sharma, 2017). Tongco (2007) asserts that in purposive sampling, the researcher needs to explicitly state what needs to be known and identify people who can provide such information. This research study aimed to collect IK indicators for malaria outbreak prediction. Therefore, the study focused on individuals with indigenous knowledge to forecast malaria outbreaks, primarily farmers and elderly people in Vhembe.

3.3.6 Data Collection

Two main approaches to data gathering are primary and secondary (Thomas, 2017). Primary data involves collecting information directly from sources, while secondary data extracts pre-existing data (Kumar, 2019). The choice of the method(s) depends

mainly on the purpose of the study and on other factors such as the availability of the resources (such as time and money) and the skills of the researcher (Casey, 2006). The researcher needs to use appropriate methods to make the respondents feel comfortable and allow data collection to be as accurate as possible (Kumar, 2019). Kumar further emphasises the critical importance of appropriate data collection methods in the research by referring to this famous old saying used in computers: "*Garbage in, garbage out*". This is because the data collecting methods provide the data input to the research, impacting the validity and reliability of the research output, the findings, and the conclusions. In this study, both primary and secondary data collection approaches were utilised.

3.3.6.1 Collecting Weather Data

Previous studies have used climate data to predict the occurrence of malaria incidences (Masinde, 2020; Segun *et al.*, 2020; Hussain-Alkhateeb *et al.*, 2021). In this research study, we obtained rainfall, relative humidity and temperature data on the Vhembe District in the Limpopo province of South Africa. The data formed part of the database used to develop the predictive models. The data time span was 1998 to 2018. The weather data were acquired from a secondary source: *the South Africa Weather Service (SAWS)* (<https://www.weathersa.co.za/>). The researcher contacted the SAWS research manager via email to request access to the data. The researcher was taken through the ethical procedure as per SAWS policies. The procedure included completing a disclosure form. The data was provided in an Excel file. The hydrological drought indices were calculated from the rainfall figures.

3.3.6.2 Collecting Malaria Data

For this study, secondary data on malaria was essential to construct the database for developing predictive models. The malaria data were obtained from the Department of Health in Limpopo Province via the South African Weather Service (SAWS). As this research is part of a larger project, the Integrated Climate-driven Multi-hazard Early Warning System for South Africa (ICMEWS) (<https://urida.co.za/icmhews-1>, [https://www.weathersa.co.za/Documents/Corporate/Media_Invite - ICMHEWS project \(Vhembe District Municipality\) 08 May 2024 15052024110957.pdf](https://www.weathersa.co.za/Documents/Corporate/Media_Invite_-_ICMHEWS_project_(Vhembe_District_Municipality)_08_May_2024_15052024110957.pdf)) managed by SAWS, the researcher was granted access to pre-existing malaria data collected by SAWS from the Department of Health in Limpopo. This dataset included structured information on monthly malaria incidences from 1998 to

2018 (21 years) in the Vhembe district. The data were anonymised, containing only the count of monthly reported malaria cases, and were provided in an Excel file format.

3.3.6.3 Collecting IK data

Indigenous knowledge data were collected to form part of the database for developing machine learning models for malaria forecasts. The data were collected from the participants from the Vhembe community in two stages using the semi-structured questionnaires and survey mobile application, respectively.

Assembling and Training of Data Collection Team

In the first stage of IK data collection, the data collection team was formed. Collecting data in a rural setting like Vhembe presents its challenges. Some participants in such rural settings may not be literate enough to understand the questionnaires and instructions. Therefore, for this data collection, we assembled a team of assistants who were familiar with the area and conversant in the local languages of the Vhembe District. These assistants were all students at the CUT and identified Vhembe district as their place of residence. As the team leader, the researcher trained and gave the team a comprehensive overview of the data collection tools and ethical considerations.

Assembling local leaders and Identifying Participants in the Vhembe Community

The researcher and the team visited the study area to facilitate conducting the questionnaires that focused on collecting IK indicators for malaria outbreaks and validating such indicators. Given the rural African context in which the study was conducted, we adopted a culturally sensitive recruitment approach rather than relying on written materials or a formal, Western-style methodology. The research leader was aware that when interacting with such rural communities as an outsider, certain norms and ethical and cultural approaches must be adhered to. Through the data collection team, we identified seven community leaders (respected custodians of community knowledge) through physical meetings conducted in adherence to local cultural norms. to identify community individuals who are well known to use and have indigenous knowledge. This contact and request to community leaders was also necessary as a sign of respect. Language translations and verbal communication were used to ensure clarity and inclusivity during these interactions. This approach was chosen to honour

the cultural context and to facilitate meaningful participation from individuals whose expertise is rooted in lived experiences rather than formal education. While this methodology differs from traditional Western recruitment techniques, it proved effective in identifying knowledgeable participants and fostering trust within the community.

From this initial group of participants, we formed a network of other potential participants on the advice of the initial group via word of mouth, social media and telephone contacts. After identifying potential participants, the research team engaged with the communities, actively reaching out to these individuals and aiding in completing the questionnaire.

Questionnaire: Collecting IK data in the Vhembe community

A questionnaire is a research instrument consisting of a series of questions to gather information from respondents (Roopa and Rani, 2012). In a questionnaire, different formats of questions and answers can be used and the researcher should choose the one appropriate for the type of question at hand (Preece, Rogers and Sharp, 2015). The use of a semi-structured questionnaire in this research study was appropriate for collecting participants' biodata, data on mobile ownership and usage, opinions of the participants on their use of IK and assessment of the participants' level of confidence in IK. The IK indicators that participants used to forecast malaria were also collected.

The researcher adopted the questionnaire previously used and validated in a study of a similar nature (Akanbi, 2019). The questionnaire (Annexure D) was piloted with ten participants before digitising it on the Survey Monkey Application. The total number of participants who completed the questionnaire was 155. The participants accessed Survey Monkey using their mobile devices through a shared link on WhatsApp. The collected IK data were later extracted into a Microsoft Excel file for analysis and the IK indicators were further stored in the database. The next chapter further discusses the questionnaire (Malaria Outbreak Knowledge Risk).

Survey Mobile Application: Collecting Observed IK indicators

In the second phase of IK data collection, a survey mobile application was employed to gather observed IK indicators remotely. Based on the first data collection phase, the researcher identified ten participants with extensive IK. These participants, referred to

as IK experts or ambassadors, were tasked with reporting IK indicators they physically observed in their environment. To facilitate this process, the IK ambassadors selected IK indicators from a database. These IK indicators were collected during the first phase of IK data collection. Before assigning the ambassadors the task of reporting IK indicators, the research and data collection team made a second visit to the study area to conduct a workshop, training the ambassadors on how to use the mobile application for reporting observed IK indicators. During this visit, the validation of IK indicators was conducted.

A mobile application was developed as part of this research to report observed IK indicators. This Android-based application, the MOEWS App, was created using Android Studio and the Java programming language. The application featured a user-friendly interface, enabling users to report text, images and location-based information.

For several reasons, the mobile app was deemed suitable for reporting observed IK indicators for malaria outbreaks. First, it allows for remote data collection, enabling the researcher to collect data without needing to be physically present at a specific location. Additionally, the app ensures that only IK indicators observed in the study area are reported, as it captures the Global Positioning System (GPS) coordinates of the reporter or IK experts. The data are collected in real-time and stored on a cloud platform, facilitating easier data analysis. The app also enables precise data reporting through data validation, allowing IK experts to accurately report IK indicators by selecting from a list constructed from the questionnaires. Furthermore, the app supports the reporting of rich data by using text, pictures and GPS coordinates, which aid in comprehensive data analysis.

3.3.7 Data Types

Data collected to answer research questions may be structured, semi-structured or unstructured (Markham, Kowolenko and Michaelis, 2015). This research study incorporated heterogeneous data to predict malaria outbreaks. The data on historical malaria incidences and weather data were structured, while IK indicators data were highly unstructured. IK indicators were documented, formatted and stored in the cloud to be used in the analysis to achieve the research objectives and answer the research questions.

3.3.8 Data Analysis and Interpretation

The data analysis and interpretation were informed by the Cross-Industry Standard Process (CRISP) for Data Mining approach (Studer *et al.*, 2021). The CRISP components considered in this research study were data understanding, pre-processing, modelling, evaluation and deployment (Studer *et al.*, 2021).

3.3.8.1 Data Understanding and Pre-Processing

In this study, data understanding was achieved by examining the dataset to identify features and the target variables and visualising the variables. Additionally, the dataset underwent pre-processing procedures to ensure data reliability, aiding the researcher in achieving the research objectives. Various techniques were applied to prepare weather and malaria data for ML modelling and predicting malaria outbreaks. Given the heterogeneous nature of the dataset, pre-processing was imperative to facilitate data integration. The weather data were recorded hourly, while malaria data were presented as monthly totals; thus, integration was necessary to align variables temporally for analysis. The monthly rainfall totals plus mean temperatures and relative humidity were computed using Python Programming Language (PPL) to combine the hourly-based weather data with the monthly malaria datasets.

The Effective Drought Index (EDI) was calculated using several parameters (Masinde, 2012; Deo *et al.*, 2017).

Step 1: Daily effective precipitation (EP) was calculated. EP is the total precipitation (P) over time, which accounts for both current and past rainfall as time progresses (Malik and Kumar, 2020). Equation 3.1 shows the daily EP formula.

$$EP = P_1 + \frac{P_1 + P_1}{2} + \frac{P_1 + P_2 + P_3}{3} + \dots + \frac{P_1 + P_2 + \dots + P_{365}}{365} \quad \text{Equation 3.1}$$

where P represents daily precipitation.

Step 2: The deviation of EP (DEP) is determined by subtracting the mean of EP (MEP) from EP, indicating the deficit or surplus of water quantity from the climatological mean.

$$EDEP = EP - MEP \quad \text{Equation 3.2}$$

Step 4: The standardised value of DEP (SEP) was computed by dividing DEP by the standard deviation of EP, providing a normalised measure of water deficit. The consecutive days of negative SEP indicate the duration of precipitation deficit, whereas consecutive days of positive SEP reflect the length of time precipitation has

been in surplus. A return to normal (PRN) determines the precipitation needed to return to normal conditions. PRN was calculated using Equation 3.3:

$$PRN_j = \frac{DEP_j}{\sum_{N=1}^j \left(\frac{1}{N}\right)} \quad \text{Equation 3.3}$$

whereby j represents the total duration, which is the number of days over which the precipitation deficit or surplus accumulates. The summation refers to the sum of the reciprocals of all months within the duration n.

Finally, EDI was computed using Equation 3.4 below:

$$EDI_j = \frac{PRN_j}{SD(DEP_j)} \quad \text{Equation 3.4}$$

SD(DEP) is the standard deviation of DEP, and EDI reflects a standardised deficit or surplus in the amount of water.

EDI values were interpreted using Table 4.1 (Byun and Wilhite, 1999; Deo *et al.*, 2017)

Table 4. 1: EDI values and their corresponding interpretation.

EDI value	Interpretation
EDI > 2	Extreme Flood
1.5 < EDI < 1.99	Severe Flood
1 < EDI < 1.49	Moderate Flood
0.01 < EDI < 0.99	Wet-Near Normal
-0.99 < EDI < 0.00	Drought-Near Normal
-1 < EDI < -1.49	Moderate Drought
-1.5 < EDI < -1.99	Severe Drought
EDI < -2	Extreme Drought

AWRI was computed using Equation 3.5.

$$AWRI = \frac{EP}{\sum_{N=1}^j \left(\frac{1}{N}\right)} \quad \text{Equation 3.5}$$

The EDI and AWRI drought indices were selected because EDI quantifies droughts and floods in absolute terms (Byun and Wilhite, 1999; Lu, 2009). AWRI's advantage lies in its ability to assess the current levels of available water resources more accurately as it measures the daily water deficit or surplus in water bodies (Wilhite and Glantz, 1985).

Data pre-processing also involved patching missing weather data points, handling outliers and normalising the data. This was achieved using PPL. Furthermore, feature selection was conducted on the dataset to pinpoint the importance of each variable in

prediction. This process enhances interpretability and focuses on pertinent variables before modelling the data. Oversampling was used in classification to handle imbalanced datasets. These data pre-processing techniques were crucial as they impacted the performance of the ML models. Additionally, the Statistical Package for Social Sciences (SPSS) was utilised as a pre-processing tool to enable qualitative analysis and visualisation of IK indicator data.

3.3.8.2 Data analysis

The structured historical weather, hydrological drought indices and malaria data were analysed using Machine Learning (ML) and Python Programming Language (PPL). Descriptive analysis was performed using Python code to determine the data pattern. Various ML models were developed using PPL. The models underwent training and testing processes to ascertain the correlations between variations in weather patterns and historical malaria incidences and make malaria outbreaks forecast. Training the ML model was necessary to allow the models to find the pattern in the dataset and make broader generalisations. In contrast, testing was crucial to ensure that models can accurately predict an unseen dataset. Both regression and classification ML models were explored.

IK data collected from the respondents in the study area were analysed using SPSS and ML. To analyse the biodata of the respondents and assess their use and level of IK, descriptive analysis was conducted in SPSS. Moreover, SPSS was used to perform a non-parametric Pearson chi-squared test and Cramer's V test to determine the level of confidence the respondents had in general use of IK, in farming and in malaria prediction, according to gender, age and migration history. These results are reported in Chapter 4.

ML models and Fuzzy Cognitive Maps (FCMs) were utilised to model IK related to malaria, contributing to the forecasting of malaria outbreaks. Firstly, the IK indicators were restructured to form concepts, which is the format required in the FCM. Causal effects of IK indicators on malaria outbreak, as stated by participants individually, were represented statistically using percentages in terms of percentage for each interacting concept. The mean of the responses was used to aggregate the participants' perceptions and determine the weights among the concepts. Consequently, weighted $n \times n$ adjacency matrices, were created to statistically represent the causal effects

among the concepts and malaria outbreak. The matrices were distinguished by four weather seasons. The FCM model was developed using concepts and their weights utilising a software package called FCM Expert. The importance of the concepts was determined using mental modeller software. The observed IK indicators collected using the MOEWS app were used to form an input to the FCM model to predict malaria outbreaks.

3.3.8.3 Error Analysis

This research study aimed to identify ML models that exhibit optimal performance in predicting malaria outbreaks. This was achieved using common performance metrics in predictions. For regression, the following metrics were used: Mean Absolute Error (MAE), Mean Square Error (MSE), Root Mean Square Error (RMSE) and R-squared (R^2), which quantify the errors and prediction accuracy of the models (Theocharides *et al.*, 2018). These performance metrics determine ML model prediction performance based on the difference between the predicted values and the actual values (Engelbrecht, 2007). A regression model is more accurate when MAE, MSE and RMSE have lower values or, equivalently, when R^2 is close to its maximum of 1 (Theocharides *et al.*, 2018). For classification, accuracy, precision, recall and F1 score (a balanced measure of precision and recall) were utilised. High values of these metrics indicate strong model performance. The formulae of these performance metrics are given in Chapter 2.

3.3.9 Reliability and Validity

Even though (Nunamaker, Chen and Purdin (1990) argue there is no standard definition of validity, in this study validity assesses the credibility of the research findings, while reliability is understood as the extent to which the method of data collection or analysis procedures yield consistent findings (Saunders, Lewis and Thornhill, 2009). Reliability and validity ensure that the results are correct and can be reproduced using the same methodology. As mentioned in Section 3.3.6.3, the researcher adopted a previously validated questionnaire and piloted the questionnaire to ensure the validity and reliability of the questionnaire's content. According to Saunders, Lewis and Thornhill (2009), a minimum of ten participants should be included in the pilot study for small-scale questionnaires. For this reason and because of the time and financial constraints, senior researchers from CUT and SAWS also

reviewed the adopted questionnaire. Consequently, minor necessary amendments were made to the questionnaire based on feedback from the pilot study and senior researchers. Furthermore, the researcher ensured the accurate and truthful collection and documentation of IK indicators on malaria prediction, by engaging with the indigenous ‘experts’ and followed ethical guidelines. To validate the IK indicators, the researcher cross-referenced the collected information with existing literature and made follow up with the participants.

3.3.10 Summary Research Design and Methodology

This section provided an overview of the research methodology and design employed in this study. To fulfil the goal of this research study, which aimed to predict malaria outbreaks utilising big climate data, IK and ML, a mixed methods research design was implemented. This involved deploying quantitative and qualitative research approaches to ensure comprehensive and effective investigation. The climate data, historical malaria incidence and IK indicator data were extracted from secondary sources while current IK indicator data was collected from primary sources using the questionnaire. Purposive and snowball sampling were used to select IK experts to ensure the target sample had the required knowledge. Data pre-processing and analysis methods, the study area and the ethical considerations for the research study were also discussed.

3.4 Chapter Conclusion

This chapter has laid out the foundational elements of the study by presenting both the theoretical framework and the detailed methodology used to investigate malaria outbreak prediction through the Malaria Outbreak Early Warning System (MOEWS). The theoretical framework for this research is rooted in the EWS architecture derived from the United Nations International Strategy on Disaster Reduction (UNISDR). MOEWS adapts the EWS components to focus on malaria, incorporating local IK alongside conventional weather data.

The chapter also detailed the research design, which aligns with the MOEWS framework. The research design employs a mixed methods approach to capitalise on the complementary strengths of utilising both qualitative and quantitative data. This design is crucial for integrating diverse data sources into a predictive model, including IK indicators, weather data and historical malaria records. The methodology involved

rigorous data collection and pre-processing, employing machine learning models to analyse and predict malaria outbreaks.

This chapter has laid a foundation for the upcoming chapters that illustrate the implementation of the three components of MOEWS, thereby providing a comprehensive understanding of how the system operates in predicting malaria outbreaks.

4 Chapter Four: Infectious Disease Outbreak Risk Knowledge

4.1 Introduction

The preceding chapter provided the research methodology and the framework for implementing an early warning system for infectious diseases. This chapter expounds upon the implementation of the first component of the framework, which deals with infectious disease outbreak risk knowledge. For this component, this study uses heterogeneous data from three distinct domains: 1. Indigenous knowledge, 2. Conventional weather station data and 3. Health data on historical malaria incidences.

Indigenous knowledge indicators (IKIs) for infectious disease prediction may differ from one region to another. Moreover, people may possess different levels of indigenous knowledge (IK) based on their migration history, age, major economic activity they engage in and level of education, among other factors. This chapter will explore these factors in the context of Vhembe district.

This study focused on the IKIs used by the community of Vhembe in Limpopo province in South Africa. As stipulated in Chapter 3, the researcher and the facilitating team used questionnaires to identify which IKI participants in the Vhembe community to use to predict malaria. The collected data were used as the basis of the IK collection, processing and dissemination design.

During the data collection of IKIs, the author, guided by the literature review in Chapter 2 and the questionnaire responses, identified and documented the unstructured malaria IKIs into seven categories. The seven categories are as follows: 1) behaviours of insects, 2) behaviours of birds, 3) plants or trees, 4) behaviours of animals, 5) astronomical and meteorological, 6) Knowledge of seasons and 7) other indicators. The latter refers to myths and religious beliefs, such as the belief that an extreme drought is a curse for offences to the gods.

The structured conventional weather data was sourced from South Africa Weather Services (SAWS) as secondary data. The meteorological and hydrological parameters that were critical for this study included precipitation, rainfall, temperature and relative humidity, effective drought index (EDI) and available water resource index (AWRI). Hydrological drought indices were computed from precipitation. The other secondary

data was the structured historical malaria incidences from the Department of Health in Limpopo Province, South Africa.

The chapter is organised into four main sections: Section 4.1 an introduction, followed by Sections 4.2, 4.3, and 4.4, which discuss the three data sources: Indigenous knowledge, conventional weather data and structured health data, respectively. The chapter concludes with a summary in Section 4.5.

4.2 Data Source 1: Indigenous Knowledge

4.2.1 Sample Data

The IK indicators relating to malaria prediction were collected between the 1st of October 2020 and 30th July 2021. This data was collected using a semi-structured questionnaire as indicated in Chapter 3.

The questionnaire was administered to 155 respondents and 146 responses were considered after data cleaning. The questionnaire consisted of four sections (see Appendix D). Section A was the introduction, highlighting the study's purpose and outlining the questionnaire's structure. Sections B and C concerned the consent form and demographic information, respectively. These sections were followed by Section D on knowledge and use of weather information and IK. The last section, Section E, focused on collecting IK indicators.

The semi-structured questionnaire aimed at acquiring the following information from the respondents:

- Biodata. Of importance from this data were the duration the respondent's stay in their villages, the main economic activities in the village and their ages.
- Knowledge of indigenous forecast to assess use of IKIs and their confidence in the reliability of IKIs.
- Mobile ownership and usage to assess the feasibility and suitability of using mobile phones to collect IKIs and disseminate infectious disease outbreak alerts.
- Knowledge of weather forecasting to assess respondents' use of weather and their confidence in that information.
- To assess the awareness of weather-related services and alerts offered to the Vhembe community.

- To collect IK indicators that the Vhembe community uses to predict malaria.
- The relevance of the IKIs to the prediction of malaria in Vhembe.

4.2.2 Data Presentation, Pre-Analysis, and Discussion

4.2.2.1 *Demographics of the Respondents*

i) **Participants' Geographical Distribution, economic activity, gender, age and migration history.**

Respondents were drawn from different villages in the Vhembe district, with most respondents being from the Tshino, Tshivhulana and Musina villages. Small scale farming, business and industry services were identified as the main economic activities. Regarding gender distribution, there were more males (67.12%) than females (32.88%) out of the 146 participants.

Information on the migration history of the respondents was also collected and the following options were given: 5-10 years at their current village location, 11-20 years and above 20 years. Over 80% of the respondents have stayed in their current village for over two decades, while 13.01% and 6.85% have stayed for 11 to 20 years and 5 to 10 years, respectively. This may indicate that the sample consisted of people who spent significant time observing and understanding the environment and possibly having rich IK for malaria prediction. Observation and understanding of indicators would seem to be associated with frequent and prolonged contact with the environment. Figure 4.1 below indicates the participants' migration history by age and gender.

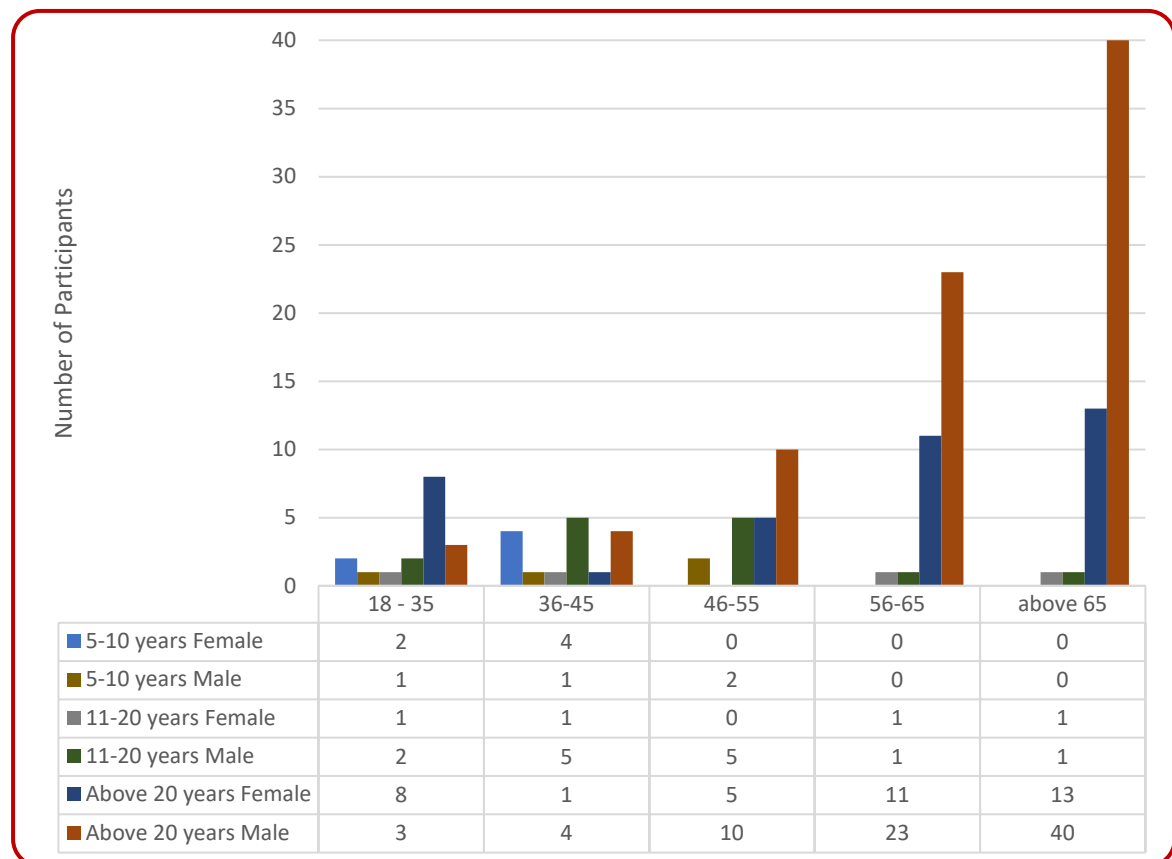


Figure 4. 1: Participants distribution by migration history, age, and gender.

The legends in Figure 4.1 indicate migration history by gender; for example, 5-10 years female means female participants who have stayed for 5-10 years in Vhembe, while the heights indicate the number of participants in each category who stayed certain numbers of years in Vhembe. It can be noted that the participants who have stayed the longest (above 20 years) in Vhembe are mostly males above 65 years old.

ii) Respondents' level of education by age and gender

Most of the respondents were elderly people above 65 years, followed by respondents in the age bracket 56-65, with these age groups representing 37.67% and 24.66% of the total sample size, respectively. The least represented group was in the age bracket 18 to 25 years, with only 11.64%. On the other hand, 17 respondents, representing 11.64% of the respondents, fell in the age bracket of 18 to 25. Respondents below the age of 18 were not represented in the data collection. The absence of respondents below the age of 18 is attributed to the purposive sampling used which aimed at including only the respondents with some IK on the

prediction of malaria. Most of the respondents (92%) showed some level of formal education, as compared to those who reported no formal education (8%). Of all participants, 64% had a post-secondary school qualification, 21% had a secondary qualification and 7% had a primary qualification.

It can also be observed from Figure 4.2 below that the participants without formal education or a low level of formal education (primary education) were elderly participants and mostly females. The number of participants with no or a low level of formal education dropped with a decrease in age, all young participants had some basic level of formal education.

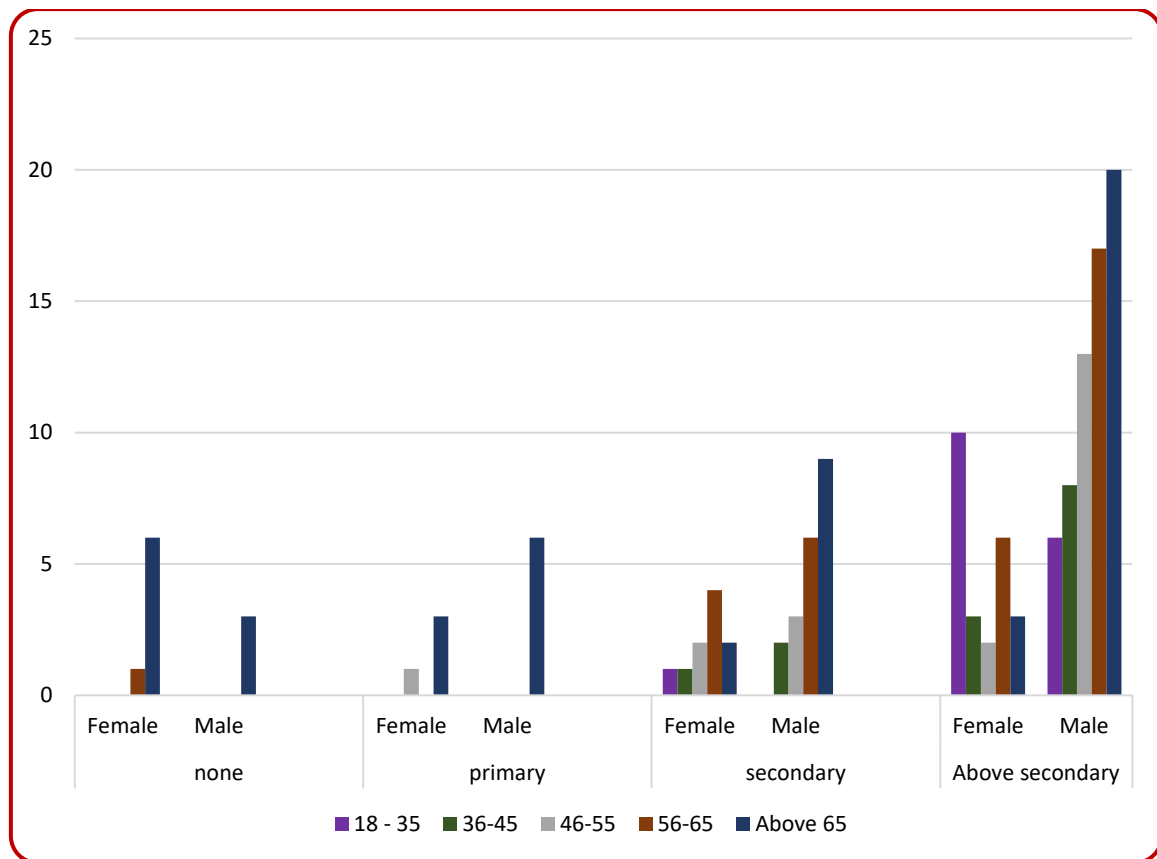


Figure 4. 2: Respondents' level of education by age and gender.

iii) Respondents' Distribution by Mobile Phone Ownership and Usage

Table 4. 2 Participants' distribution by Mobile ownership and Usage.

	percentage	frequency
Yes, I own a phone	97.95%	143
No, I do not own a phone	1.37%	2
I do not own a phone, but I have access to a phone	0.68%	1

As indicated in Table 4.1, almost the entire sample (97.95%) owned a mobile phone except for three respondents (1.37%), of whom two neither owned phones nor had access to a phone, and one respondent (0.68%) did not own a phone but had access to a phone. Most participants (81.94%) owned smart mobile phones while the remainder (18.06%) owned phones but not smartphones. This high mobile phone penetration among the Vhembe community encourages the exploration of mobile phones to collect IKIs and disseminate malaria outbreak alerts.

iv) Respondents' knowledge of weather forecast information.

The respondents were also asked about the use of weather information in the Vhembe community and the sources they used to acquire this information. The options were radio, TV, Internet, newspapers/print media, website/App, local observation (IKIs) and others. Figure 4.3 illustrates the frequency of checking weather information by respondents.

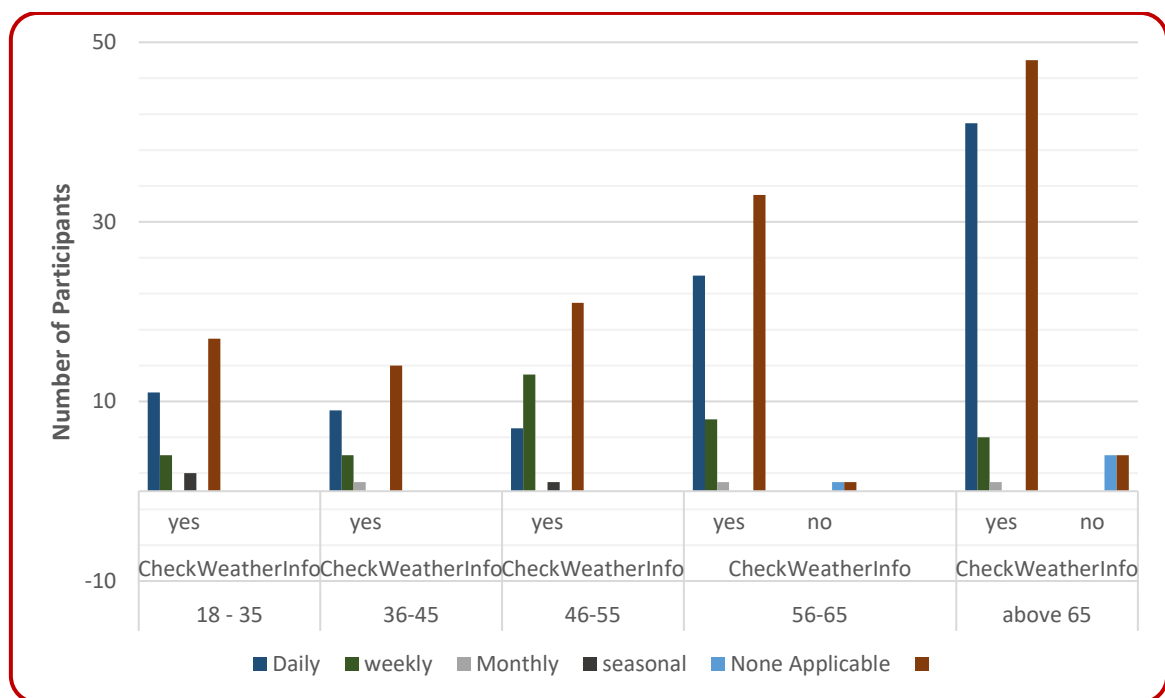


Figure 4. 3: Frequency of checking weather information by Age.

Most of the respondents, 95.65%, indicated that they have been using weather information for over five years, with 69,17% checking weather information daily, 26.32% checking it weekly and 2.26% checking it monthly and seasonally. The source of weather information mostly used was television (TV) followed by radio.

Other participants, 95, 68 and 105, checked the weather information from websites/Apps, newspapers and using IKIs, respectively.

The respondents were also asked to evaluate sources of weather information and to indicate the relevance of the weather information to their needs and their confidence in such information sources. Number 1 indicates the least confidence while 5 indicated the highest confidence. Most respondents showed high confidence in the weather information provided on TV, radio and local indicators as indicated by Figure 4.4.

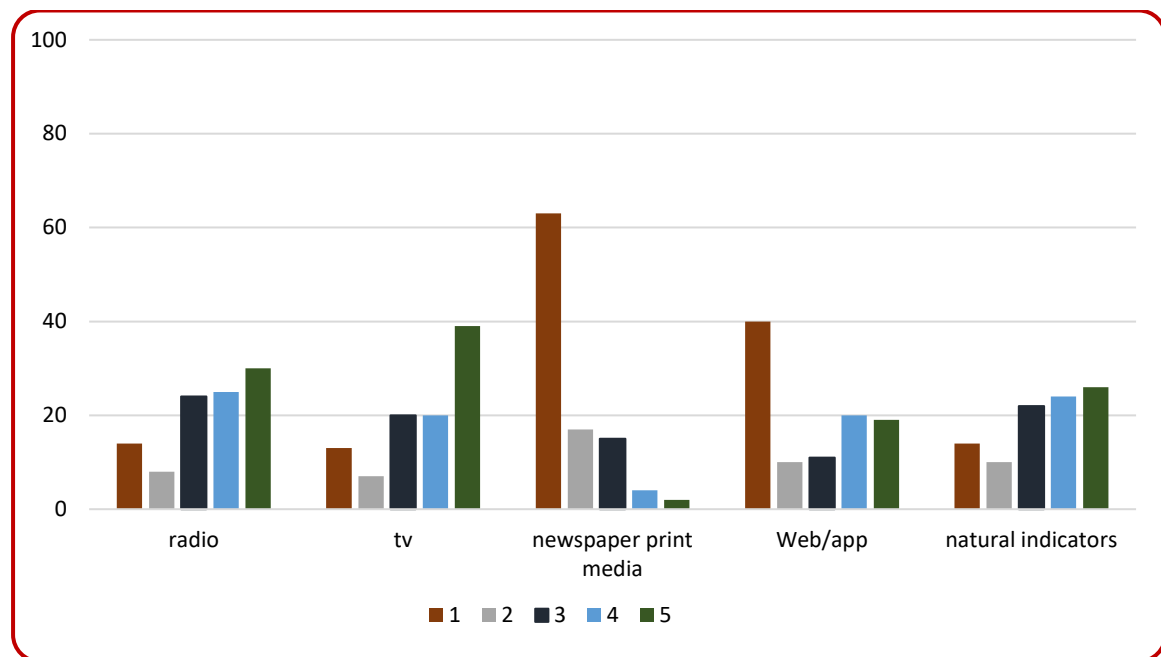


Figure 4. 4: The rating of the sources of weather information.

Most participants indicated that the weather information they receive has a low level of adequacy in predicting malaria. For farming, travelling and planning the day, most participants asserted that the weather information has a moderate level of adequacy. Those who gave a lower rating were asked to justify their ranking. Some of the reasons included that the weather information is not accurate anymore, not specific to the individual needs and not communicated clearly to the community.

v) Respondents' distribution of weather-related alerts/forecast

Participants were also asked to name the weather-related alerts they received and the communication channels they mostly used to acquire this information. While the options given for alerts were heat stress, air quality, ultraviolet, malaria and others,

for the communication channels, radio, TV, Internet, newspapers/print media, website/App, local observation (IKIs) and others were given. The participants also stated that they use weather information to inform their activities, which included farming, travelling, predicting infectious diseases and planning the day (such as when to do laundry, how to dress or when to take an umbrella). The weather information was found to be used more for farming activities than for the prediction of infectious diseases.

Most respondents (82.61%) stated that they receive alerts. Heat stress, thunderstorms and wildfires, mentioned under the 'other' category, were identified as the common alerts. Figure 4.5 depicts weather-related alerts received by participants in Vhembe and the communication channels.

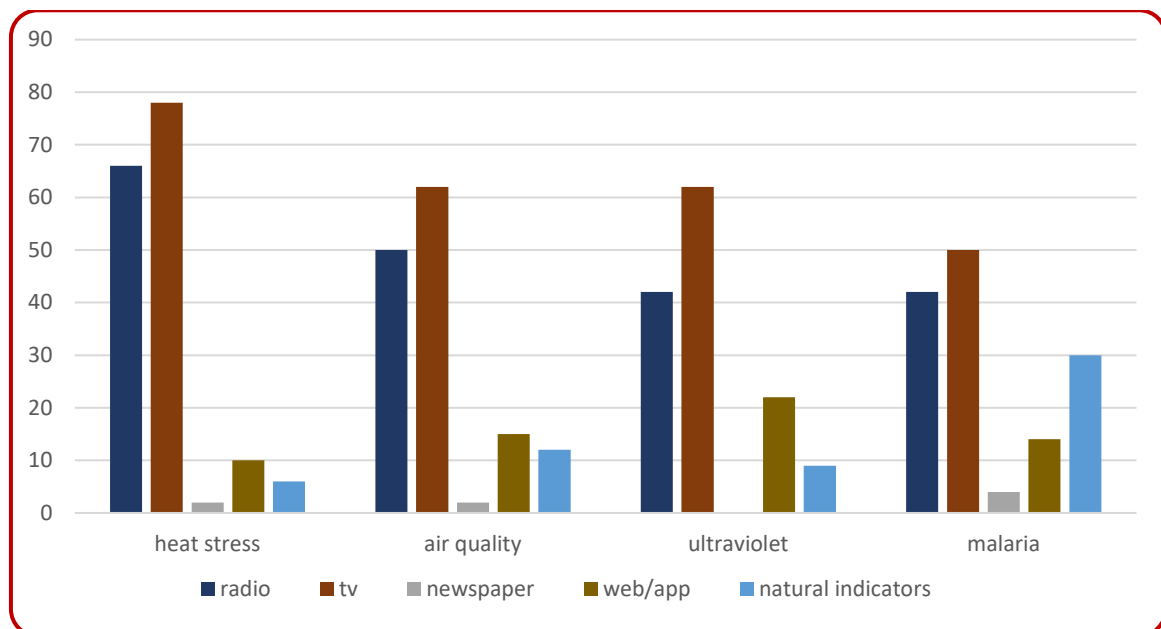


Figure 4. 5: Weather-related alerts received by participants in Vhembe and the channels of communication.

Radio and TV were the most common communication channels for acquiring the alerts, while newspaper and local indicators were the least common channels. SMS was indicated by some respondents under the other category. Moreover, the natural indicators were identified as a common source for malaria outbreak alerts.

When asked to rate the accuracy of the sources of weather forecast information, TV and radio followed by natural indicators, were rated as the most accurate sources of weather forecast information, as seen in Figure 4.6. Newspapers were the least rated source of weather forecast information.

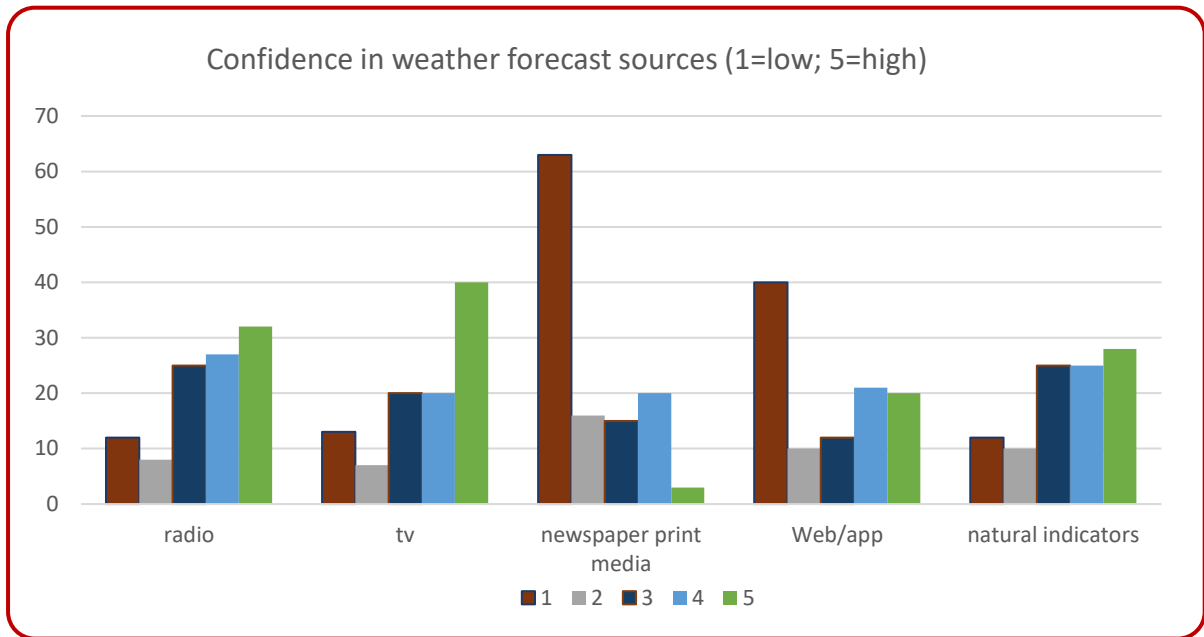


Figure 4. 6: level of confidence in the accuracy of the sources of weather forecast information.

4.2.2.2 Confidence in the Use of Indigenous Knowledge by the Vhembe Community

Inferential statistics were used to analyse the level of confidence in using IK. Of interest were the confidence in the general use of IK and the confidence in the use of IK for disease prediction and farming by gender, age and migration history. Firstly, the number of participants who use IK was determined and illustrated in Table 4.2 below.

Table 4.3: Use of IK by the Vhembe community.

	Participants who use IK	Participants who do not use IK	Total
Female	36	9	45
Male	69	24	93
Total	105	33	138

Participants possess rich IK that is used for various purposes. This is indicated by 76%, 105 of 138 participants, that use IK. The number of participants was reduced from 146 to 138 because, in the questionnaire under the questions on IK use, eight participants left the questions unanswered and these so participants were excluded.

i) Confidence of Participants in General Use of IK

In this section, analyses are conducted to determine the confidence of the participants in the general use of IK by gender, age, and migration history.

a. Confidence of Participants in General Use IK by Gender

Figure 4.7 shows the confidence of participants in the use of IK by age and gender.

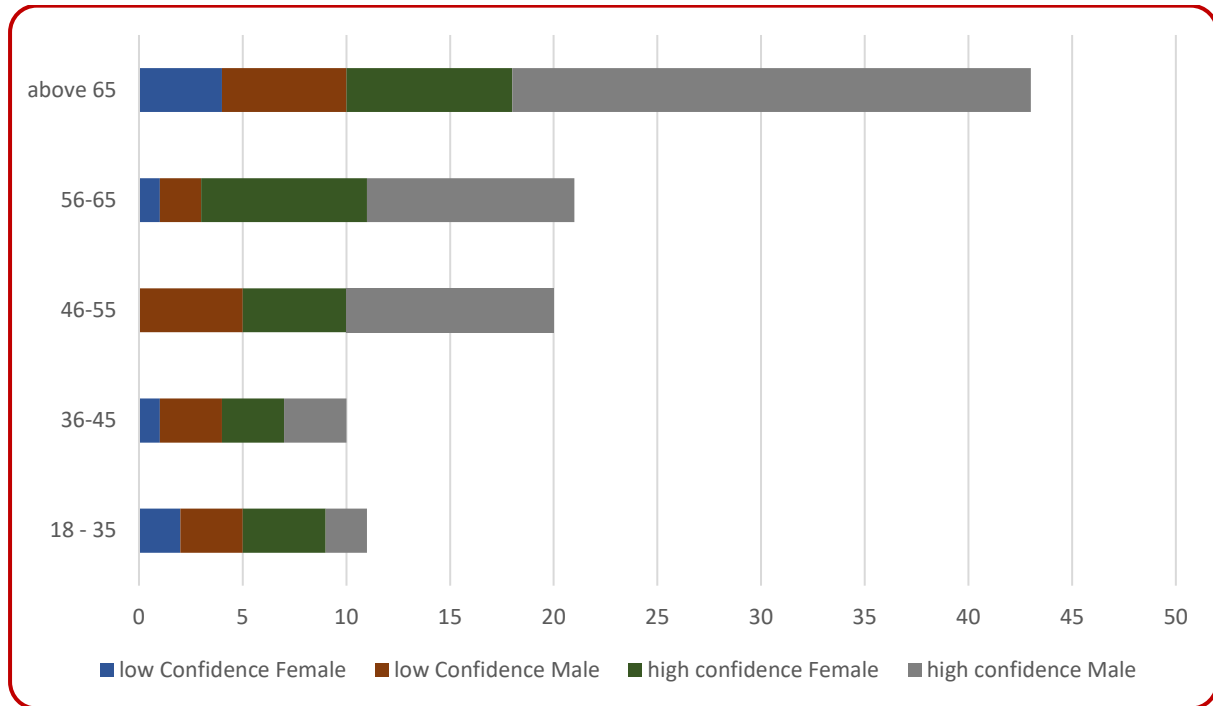


Figure 4.7: confidence in general use of IK according to gender and age.

It can be noted that elderly people have more confidence in the use of IK compared to younger participants. It can also be observed that there is a higher number of male participants who have more confidence in the general use of IK compared to their female counterparts.

To determine the level of confidence among the participants who use IK, the non-parametric Pearson Chi-squared test for independence was used on the data (not normally distributed). Firstly, the contingency table (Table 4.3 below) of the observed and expected data for confidence in the general use of IK was constructed after collapsing the Likert scale (very inadequate and inadequate into one column and, adequate and very adequate into another column).

Table 4. 4: Contingent table for confidence in general use of IK according to gender.

	Inadequate Confidence	Adequate Confidence	
Female	8	28	36
Male	19	50	69
Total	27	78	105
Theoretical f/m ratio			0.52
Observed f/m ratio	0.42	0.56	

The observed ratio is calculated directly from the given data for each confidence category. The ratio is the proportion of females to males and the calculations are displayed in Equations 4.1 and 4.2.

$$ObservedIC = \frac{females}{males} = \frac{8}{19} \approx 0.42 \quad \text{Equation 4.1}$$

$$ObservedAC = \frac{females}{males} = \frac{28}{50} \approx 0.56 \quad \text{Equation 4.2}$$

ObservedIC and ObservedAC represent observed ratios for inadequate and adequate categories, respectively.

The same approach was followed to calculate theoretical female/male (f/m) ratio which reflects the overall proportion of females to males in the dataset. The calculations are shown in Equation 4.3.

$$T = \frac{total_{females}}{total_{males}} = \frac{36}{69} \approx 0.54 \quad \text{Equation 4.3}$$

Where T is the theoretical f/m ratio.

It could be observed that most participants (74%) have confidence in the general use of IK. The theoretical and observed ratio between females and males who have used IK was calculated to determine whether there is any association between the gender of the participants and their confidence in the general use of IK.

As depicted in Table 4.3 and in the equations above, the expected female/male (f/m) ratio is 0.52 in the original sample for the general use of IK. For those choosing adequate confidence in the general use of IK, the observed f/m ratio is 0.56 (> than 0.52), therefore proportionally more females in the sample have confidence in the use of IK than male participants. Their male counterparts make less use of IK (0.42 experimental ratio compared to 0.52 theoretical ratio). This association was verified through an inferential statistics procedure for any significant difference whereby a non-parametric Pearson chi-squared test was chosen at an alpha level of 0.05.

For the Chi-Squared test, the following Null Hypothesis was formed: There is no significant difference between male and female participants in their confidence in the

use of IK. In other words, there is no association between gender and confidence in the general use of IK among the participants.

The Chi-Squared test results were as follows: $\chi^2 = 10.851$, $p = 0.554$ ($N=105$) and from this result ($P > 0.05$). We, therefore, accept the null hypothesis. We conclude that there is no significant difference in the general use of IK among the participants according to gender, even though Figure 4.7 and Table 4.3 suggest some association between the gender of the participants and their confidence in the use of IK in general.

b. Confidence of Participants in General Use of IK According to Age

Of interest also was the confidence of the participants in the general use of IK according to age. The contingency table below, Table 4.4, indicates the confidence of the participants by age.

Table 4.5: Contingency table for the confidence in general IK use according to age.

	Inadequate Confidence	Adequate Confidence	Total
18-35	5	6	11
36-45	4	6	10
46-55	5	15	20
56-65	3	18	21
Above 65	10	33	43
Total	27	78	105

Cramer's V was used where the dependent variables had more than two levels (such as age). In such instances, the Chi-squared would not be applicable (Field, 2013). Cramer's V was used to investigate the level of association of confidence in the use of IK with age. Cramer's V ranges from 0 to 1, where 0 indicates no relationship and 1 indicates perfect association (Field, 2013). Cramer's V test was run on the data to verify the association between age and confidence in the general use IK. The Cramer's V results of the confidence in general use of IK by age were as follows: Cramer's V = 0.216, $p=0.029$. The interpretation of Cramer's V results was guided by Table 4.5, which demonstrates different ranges of Cramer's V and their interpretations (Field, 2013).

Table 4.6: Cramer's V and interpretation (Akoglu, 2018).

Cramer's V	Interpretation
> 0.25	Very strong

> 0.15	Strong
> 0.10	Moderate
> 0.05	Weak
>0	No association/very weak association

With Cramer's V at 0.216, we conclude that there is a strong association in the use of IK among the participants according to age. This coincides with the results shown in Figure 4.7.

c. Confidence of participants in general use of IK by migration history

Figure 4.8 and Table 4.6 below indicate the confidence of the participants in the general use of IK by migration history. It can be observed the participants who stayed longer in Vhembe have more confidence in the general use of IK.

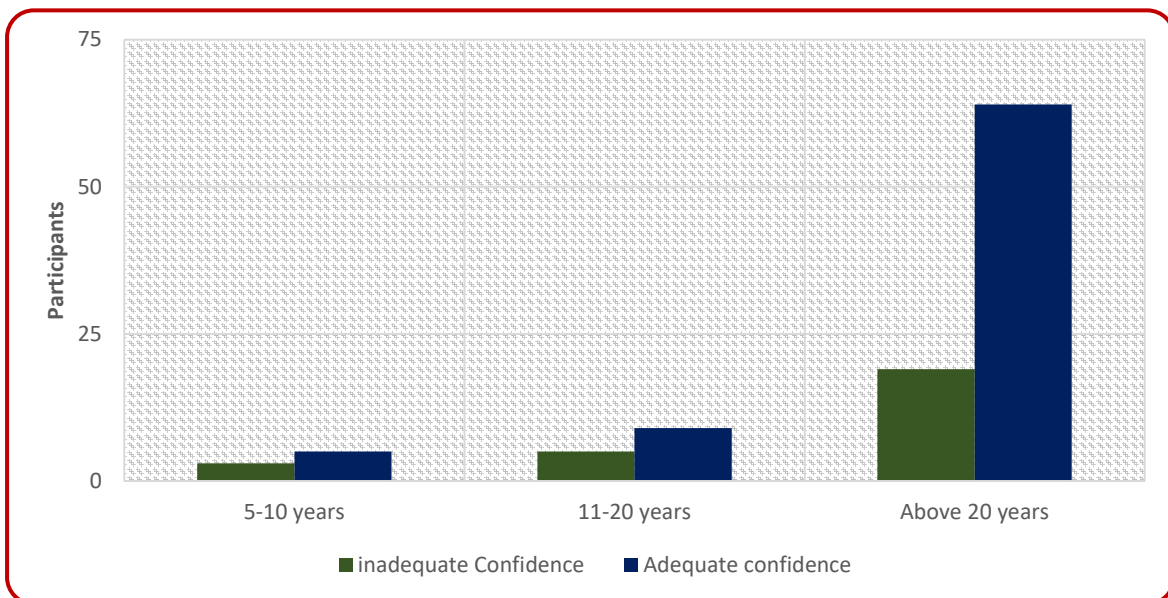


Figure 4.8: Participants' confidence in general use of IK by migration history.

Table 4. 7: Contingent table for the general use of IK by migration history.

	Inadequate Confidence	Adequate Confidence	Total
5-10 years	3	5	8
11-20 years	5	9	14
Above 20 years	19	64	83
	27	78	105

Cramer's V analyses were run on the general use of IK and migration history data to verify the association between migration history and confidence in the general use of IK. The following results were obtained: Cramer's V = 0.126, $p=0.044$. We conclude that there is a moderate association in the use of IK among the participants according to migration history.

ii) Confidence of Participants in IK Use for Disease Prediction

a. Confidence of participants in IK for disease prediction by gender

Of 105 participants who indicated to use of IK, 87 participants (82.9%) indicated that they have confidence in the use of IK for disease prediction. These participants make up 63% of the total sample, 138. Figure 4.9 below depicts the distribution of the use of IK for disease prediction by age and gender.

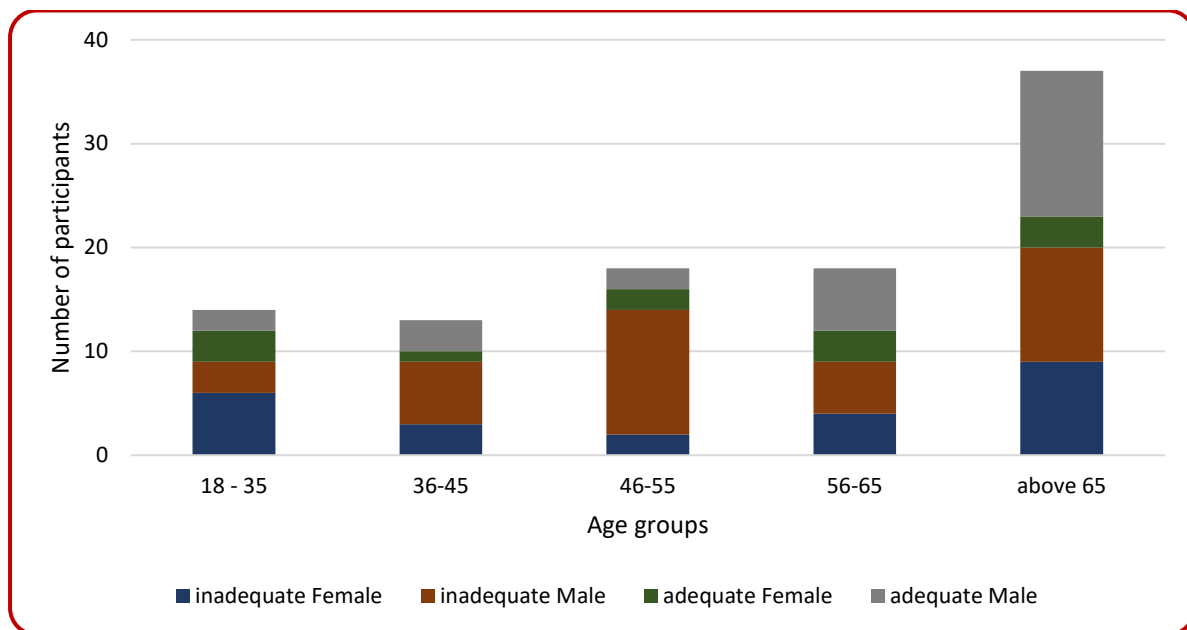


Figure 4. 9: Confidence in IK use for disease prediction by gender and age.

It can be observed from the figure above and the contingent table below that males who are 66 years old and above seem to have more confidence in the use of IK for disease prediction. Generally, male participants seem to have more confidence in the use of IK for disease prediction compared to the female participants.

Table 4.7 is a contingent table showing gender and confidence in IK use for disease predictions.

Table 4. 8: Contingent table for gender and confidence in IK use for disease prediction.

	Inadequate Confidence	Adequate Confidence	Total
Male	32	24	56
Female	18	13	31
Total	50	37	87
Theoretical m/f ratio			1.81
Observed m/f ratio	1.78	1.85	

From the contingency table above, the expected ratio is 1.81 considering the male/female sample ratio in the original sample. For those choosing inadequate confidence ratio is 1.78 (< than 1.81). Therefore, proportionally, females have less confidence in their use of IK for disease prediction than male participants. Their male counterparts feel more confident in their use of IK for disease prediction (experimental ratio of 1.85 compared to theoretical ratio of 1.81). The association was verified for significance difference using the Chi-squared test, at an alpha level of 0.05 and the following null hypothesis was formed: There is no significant difference between male and female participants in their confidence in the use of IK in disease prediction. There is no association between gender and confidence in IK use for disease prediction among participants. The Chi-Squared results were as follows: $\chi^2 = 10.851$, $p = 0.744$ (N=87). From this result, we accept the null hypothesis. The results suggest that there is no association between gender and confidence in IK use for disease prediction among participants even though the observation from the graph in Figure 4.9 and the contingent table above suggest that male participants have more confidence in their use of IK for disease prediction.

b. Confidence of participants in IK use for disease prediction by age

This section aims to determine the association if any between the age and confidence of participants in their use of IK for disease prediction. The contingency table below (Table 4.8) shows the confidence of the participants in the use of IK for disease prediction by age.

Table 4.9: Contingency table of confidence in IK use for disease prediction by age.

	Inadequate Confidence	Adequate Confidence	Total
Under 18	0	0	0
18-35	6	5	11
36-45	7	3	10
46-55	14	4	18
56-65	5	8	13
Above 65	18	17	35
Total	50	37	87

Cramer's V was run on the data to verify the level of association between age and confidence in IK use for disease prediction. The null hypothesis was as follows: there is no association between the age of participants and their confidence in IK use for disease prediction. If we reject the null hypothesis the alternative hypothesis will be accepted, that there is an association between age and confidence in IK use for disease prediction, the Cramer V calculation is carried out to determine the level of association. The following results were obtained. Cramer's $V = 0.27$, $p=0.0187$. As per Table 3, with Cramer's V at 0.27, we conclude that there is a very strong association in the use of IK among the participants according to age. As seen in Figure 4.7, participants above 66 years old, have more confidence in the use of IK for disease prediction as compared to the younger generation.

c. Confidence of participants in the IK use for the prediction of disease by migration history

Table 4.9 illustrates the confidence the Vhembe community has in the use of IK for disease prediction by migration history.

Table 4.10: Contingent table for the confidence in the use of IK for disease prediction by migration history.

	Inadequate Confidence	Adequate Confidence	Total
5-10 years	5	3	8
11-20 years	10	3	13
Above 20 years	35	31	66
Total	50	37	87

Even though it could be observed that more participants have less confidence in the IK use for disease prediction, participants who resided for more than 20 years in Vhembe have more confidence in the use of IK for disease prediction, than those who resided for a shorter time. The association between the migration history and confidence in the use of IK for disease prediction as determined in Table 4.9 was tested using Cramer's V.

The null hypothesis was as follows: there is no association between the migration history of participants and their confidence in IK use for disease prediction. If we reject the null hypothesis the alternative hypothesis will be accepted, that there is an association between migration history and confidence in IK use for disease prediction, the Cramer V calculation is carried out to determine the level of association. The results were as follows: Cramer's V = 0.17, $p=0.027$. we conclude that there is a strong association in the use of IK for disease prediction among the participants according to migration history.

iii) Confidence of Participants in the Use of IK for Farming

The results suggest that respondents used IK for farming more than for disease prediction. Out of the total sample, 104 respondents (75.3%) and 87 respondents reported using IK for farming, and disease prediction, respectively. Most respondents (82.3%) expressed adequate confidence in the use of IK for farming.

A similar procedure as above was followed to determine the confidence of participants in the use of IK for farming by gender, age, and migration history.

Based on the null hypothesis: there is no difference between male and female participants in their confidence in the use of IK in farming, and the results: $\chi^2 = 3.849$, $p = 0.94$, (N=104), it can be concluded that there was no association between gender and confidence in the use of IK for farming.

When it comes to age the following results were obtained: Cramer's V = 0.02, $p=0.039$. The results suggested that there was a very weak to no association between age and confidence in the IK use for farming.

Different results were obtained for the confidence in IK use for farming according to migration history and these results are discussed below.

Confidence of participants in the use of IK for farming by migration history

The contingency table below, Table 4.10, and Figure 4.10 below indicate the confidence participants have in the use of IK for farming according to migration history.

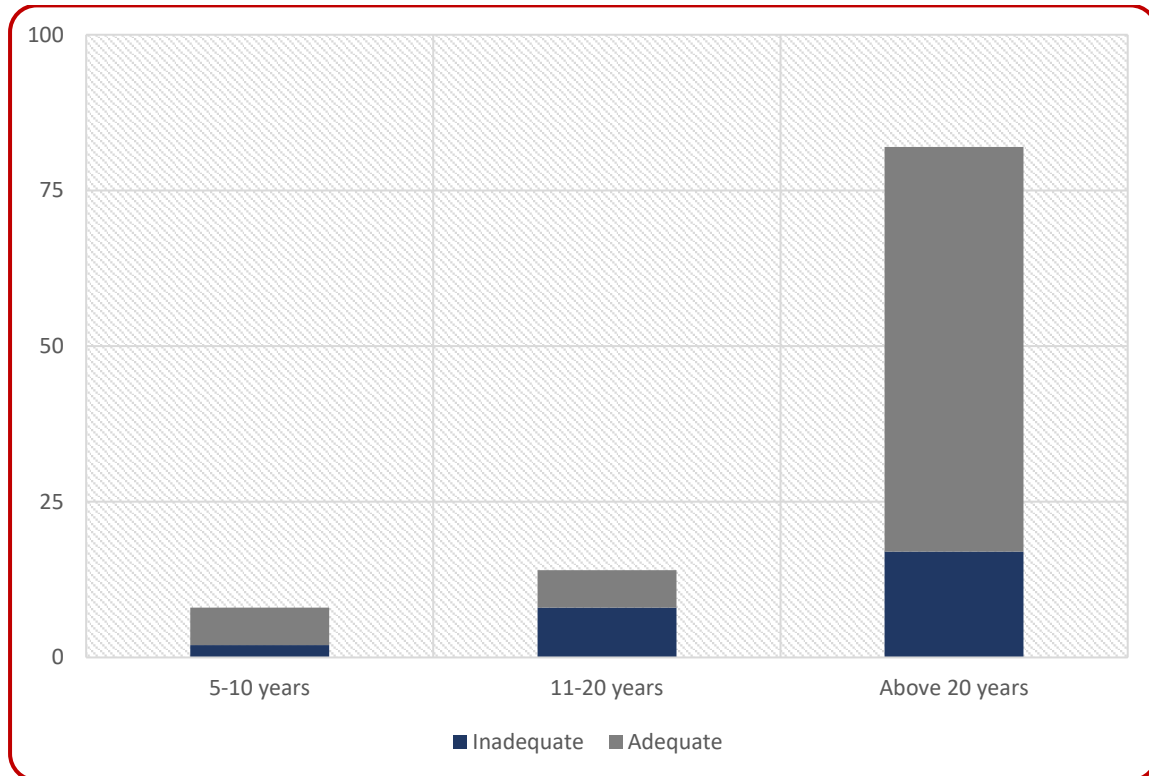


Figure 4.10: Confidence in the IK use for farming according to migration history.

It can be observed that participants with 20 years or more of staying in Vhembe have more confidence in the use of IK for farming. This association between the migration history and the confidence in IK use for farming was verified using Cramer's V test.

Table 4. 11: Contingency table for use of IK for farming by migration history.

	Inadequate Confidence	Adequate Confidence	Total
5-10 year	2	6	8
11-20 years	8	6	14
Above 20 years	17	65	82
Total	27	77	104

The following null hypothesis was formed to test the significance of the above mentioned: association: there is no association between the migration history of the participants and their confidence in the use of IK for farming. If we reject the null hypothesis the alternative hypothesis will be accepted, that there is an association

between the migration history of the participants and their confidence in IK use for farming.

As per the following results, Cramer's $V = 0.28$, $p=0.016$, there is a very strong association in the use of IK for farming among the participants according to migration history.

Table 4.11 below summarises the results of all the analyses done above to determine the confidence of the participants in the general use of IK and the use of IK for disease prediction and farming according to age, gender and migration history.

It is important to also note that participants use IK for farming more than they do for disease prediction. 104 participants used IK for farming compared to 87 who used IK for disease prediction.

Table 4.12: Summary of the confidence participants from the Vhembe community have in IK use according to gender, age, and migration history.

	Confidence in general use of IK	Confidence in the use of IK for	
		Disease prediction	Farming
Gender	No association	no association	no association
Age	Strong association (Cramer's $V=0.216$)	Very strong association (Cramer's $V = 0.27$)	Very weak to no association (Cramer's $V=0.02$)
Migration history	Moderate association (Cramer's $V = 0.126$)	Strong association (Cramer's $V = 0.17$)	Very strong association (Cramer's $V = 0.28$)

4.2.3 IK Indicators

4.2.3.1 *Characteristics of Weather Seasons in the Vhembe District*

The participants stated that there are four seasons in Vhembe: 'Tshilimo', 'Lutavula', 'Vhuria' and 'Tshifhefho'. These are the local names for summer, autumn, winter and spring, respectively. The participants indicated that through long observation and understanding of their environment, they have identified the signs that indicate the onset and cessation of each season. They also have local indicators that characterise each season. For example, the onset sign of spring is the strong wind while the cessation of spring is marked by high temperatures. Springs are characterised by the

flowering of trees and plants. Table 4.12 below gives details of the four weather seasons in Vhembe with their corresponding onset and cessation signs.

Table 4.13: Weather Seasons in Vhembe District.

Season	Local Name	Onset Signs	Cessation signs	Local Indicators	Duration	
					start	end
Summer	<i>Tshilimo</i>	Very Hot Weather Heavy Rains	Hot Winds Continuous Rains	Nemeneme, Madzhulu, Phame, Mbenzhele ants	Nov	Feb
Autumn	<i>Lutavula</i>	Trees shed leaves	Warm weather	Avocados	Mar	Apr
Winter	<i>Vhuria</i>	Cold Weather	Cold winds	Oranges, Nartjies	May	Jul
Spring	<i>Tshifhefho</i>	Lot of Winds	Hot weather	Flowers on trees	Aug	Oct

4.2.3.2 IK Indicators Used by Vhembe Participants

The identified IK holders, IK experts, elderly people and farmers in Vhembe, provided a list of IKIs for malaria prediction and their in-depth descriptions. The IKIs were collected through a questionnaire. The diversity and level of detail of the indicators from the respondents confirmed that participants in the Vhembe community have very rich IK to predict malaria. The local indicators given by the participants in Vhembe were grouped into seven categories, as seen in Figure 4.11.

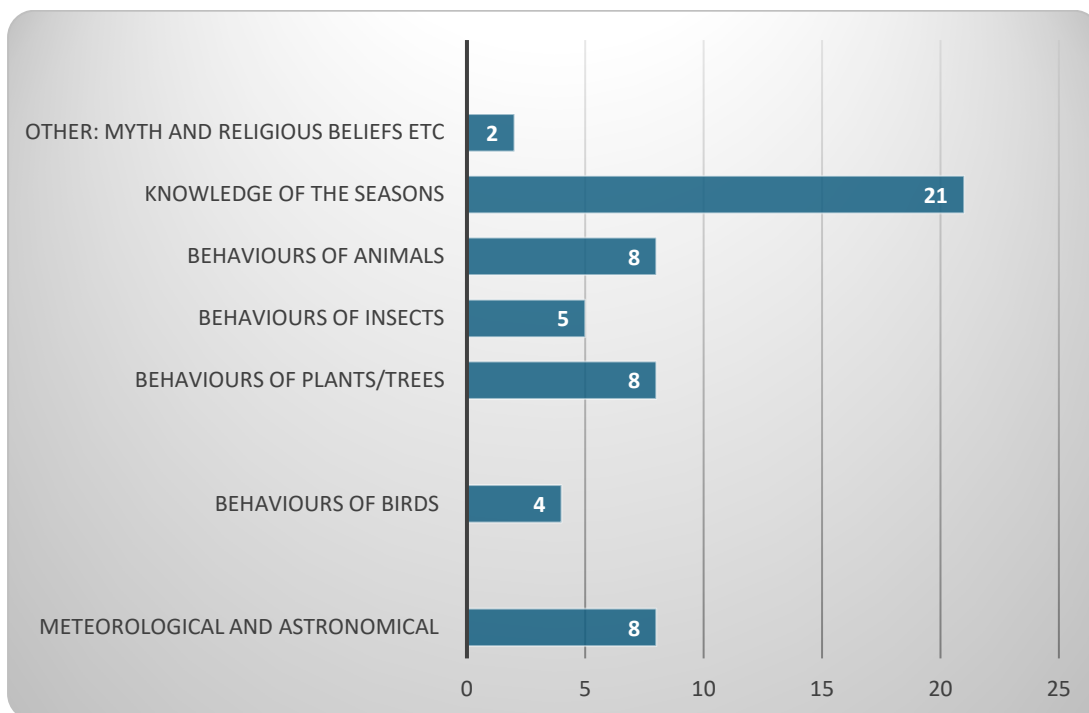


Figure 4.11: Categories of IKIs used by participants from the Vhembe Community.

When asked to specify the group of IKIs used, 37.5% claim to use knowledge of the seasons, 14.3% meteorological and astronomical behaviours of animals and the same percentage for behaviours of plants/trees each. On the other hand, 8.9% use behaviours of insects while 7.1% and 3.6% use behaviours of birds and other factors (Myth and Religious Beliefs etc), respectively. It can be noted that participants from the Vhembe community have rich IK about the seasons of the year that they use to predict malaria and rainfall/drought. It was also observed that most of the IKIs are based on the behaviours of plants/trees and animals. The least used IKIs for malaria predictions are the ones based on myths, religious beliefs and others. Table 4.13 shows the IK indicators Vhembe participants used. These indicators are grouped according to their respective categories and season.

Table 4. 14: Malaria IKIs for the Vhembe community.

	Summer (Nov-Feb)	Autumn (March-April)	Winter (May-July)	Spring (Aug-October)
Astronomical and meteorological	High temperatures during summer along with heavy rains are followed by malaria outbreaks. When the moon is tilted this means there will be a lot of rain and malaria. At the sight of the moon and the sun surrounded by a lot of clouds, floods/heavy rain and malaria can be expected. The sight of stars clustered not in the middle of the sky but along the horizon indicates long soft rains that will be followed by malaria. The moon surrounded by a ring of stars is an indication of upcoming rain and malaria season. The bigger the ring, the higher the incidences of malaria can be expected.	If it rains a lot in autumn, malaria can be expected.		If it rains in August and September, it indicates there will be plenty of rain and malaria in summer.
Behaviours of Birds	The sight of many birds indicates rainy and malaria seasons. The sight of many swallows flying around means the onset of rain and malaria season is imminent.		The sight of birds that are not usually seen means little rain and less malaria in summer.	The sight of many birds indicates rainy and malaria seasons
Behaviours of Trees/Plants	Black sports on mangos are malaria forecast. When there are too many mangos, guavas, jackal berries and marula fruit falling from the trees, malaria can be expected. Tree leaves when they lose their green colour – drought can be expected.		When fig trees “Muhuyu” do not shed their old leaves, this indicates little rain and less malaria.	Many flowers on the mango trees are an indication of upcoming malaria. When the Quinine Tree (<i>Rauvolfia caffra</i> Sond) “Munadzi” tree drips, malaria can be expected. When “mofafa grass” has many ticks, this indicates the start of rain and malaria season. When mangos trees have more flowers, and they are followed by rains and wind, there will be more malaria incidences.

Insect behaviour	When ants gather food and store them underground this signifies the imminence of the rain and malaria season. Visibility of butterflies, crows and millipedes indicates a rain forecast and malaria can be expected.	When ants gather food and store them underground signifies the imminence of the rain and malaria.	The sight of little or no insects, especially locusts, indicates that the upcoming spring and summer will have a lot of rain and malaria.	When ants gather food and store them underground, this signifies the upcoming rain season and indicates the imminence of the malaria season.
Animal Behaviour	The sight of many sardines in dams in summer indicates good rains and malaria can be expected. The hearing of frogs croaking is an indication of a rainy summer season and more malaria incidences. Frequent sight of cattle running with tails raised is a sign that they are happy for the upcoming rain season and malaria incidences can be expected.	Frequent sight of cattle running with tails raised is a sign that they are happy for the upcoming rain season and malaria incidences can be expected.		A lot of Cattle Egrets around cattle indicate rainy seasons and a lot of malaria incidences.
Myth, Religious Beliefs and others	After a period of 5–7 years, there is a lot of rainfall and malaria. After another period of 5–7 years, we experience drought and fewer malaria incidences. When heavy rains in summer stop and water on the ground recedes forming small pools of water, a malaria outbreak can be expected.			
Knowledge of Season	In summer, there is usually malaria. A lot of rain or water on land is followed by malaria. It always rains extremely in summer with high temperatures followed by malaria.	Malaria usually occurs in autumn.		If there is no rain in spring, there will be drought and fewer or no malaria incidents in summer can be expected. Rains in August and September indicate that there will be a lot of rain in summer and malaria can be expected.

IK is often subjective and context-specific, shaped by localised experiences and interpretations (Macherera, Chimbari and Mukaratirwa, 2017). During this research study, no significant conflicts or disagreements were observed among participants regarding IK Indicators and their interpretations. While disagreements in IK indicators could be expected across a large geographical area due to variations in environmental and climatic conditions (Mafongoya *et al.*, 2017), this study focused on the micro-level context of the Vhembe district, which shares a relatively homogeneous ecological and cultural environment. This shared setting likely contributed to the consistency in the IK Indicators reported by participants and the interpretation. Furthermore, the recruitment process (discussed in Section 3.3.6), which purposively selected individuals recognized for their expertise in IK by community leaders, ensured that the sample included knowledgeable and experienced respondents with a deep understanding of the indicators. The methodology also involved a workshop with IK ambassadors, identified as experts, who reviewed and confirmed the indicators. This reinforced the uniformity of the IKIs and minimised potential discrepancies.

4.3 Data source 2: Conventional Weather Station Data Collection

The weather data used in this research study was the secondary data from conventional weather stations obtained by SAWS as discussed in Chapter 3. The weather parameters considered in this study were precipitation, relative humidity, temperature, Available Water Resource Index (AWRI) and Effective Drought Indices (EDI) for a period of 21 years (1998 to 2018) from Vhembe.

4.3.1 Temperature

According to the dataset, Vhembe has cool winter temperatures and warm to very hot summer temperatures. The highest recorded temperature in Vhembe was 41⁰C and the lowest was 2.2⁰C on 11 October 2010 and 17 June 2010, respectively. Figure 4.12 shows the daily temperature in degrees Celsius from Vhembe District from 1998 to 2018.

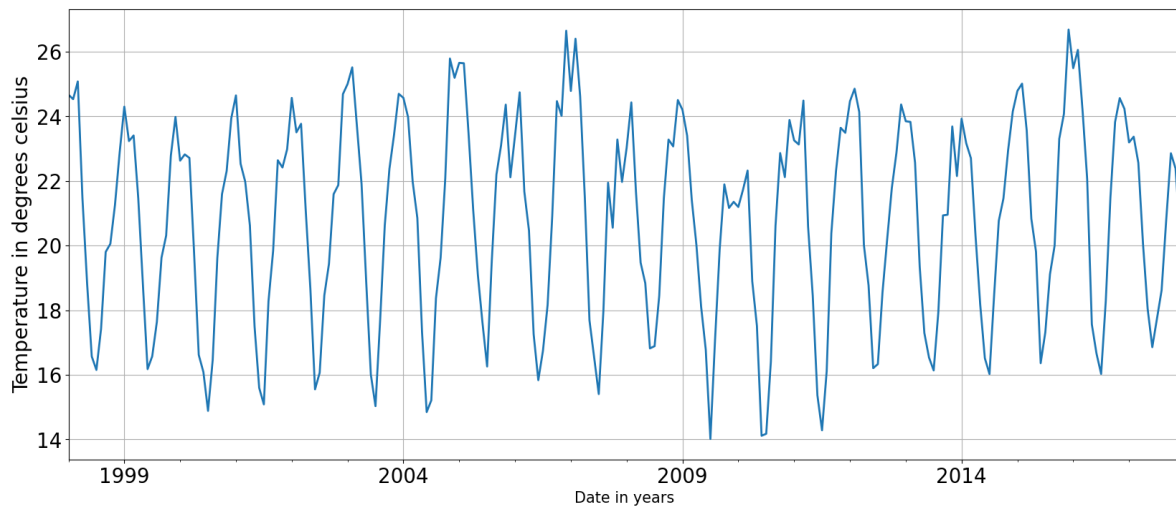


Figure 4.12: the monthly average temperature in Vhembe from year 1998 to 2018.

Observing the trend reflected in Figure 4.12, there is a clear seasonal cycle in temperature, with regular fluctuations between higher and lower values. Each year, the temperature exhibits a distinct pattern of rising to a summer peak, typically around 24°C to 26°C, before declining to a winter low, usually between 14°C and 16°C. This cyclical behaviour is consistent across the entire period from 1998 to 2018, indicating a strong seasonal influence on the region's climate. However, there are noticeable variations in the amplitude of these fluctuations. For instance, the period from 2004 to 2009 shows greater variability in temperature, with more pronounced peaks and troughs compared to other years. This increased variability could be indicative of unusual climatic events or changes to the region's climate during that time.

4.3.2 Relative Humidity

Figure 4.13 shows monthly average relative humidity recorded in Vhembe from year 1998 to 2018.

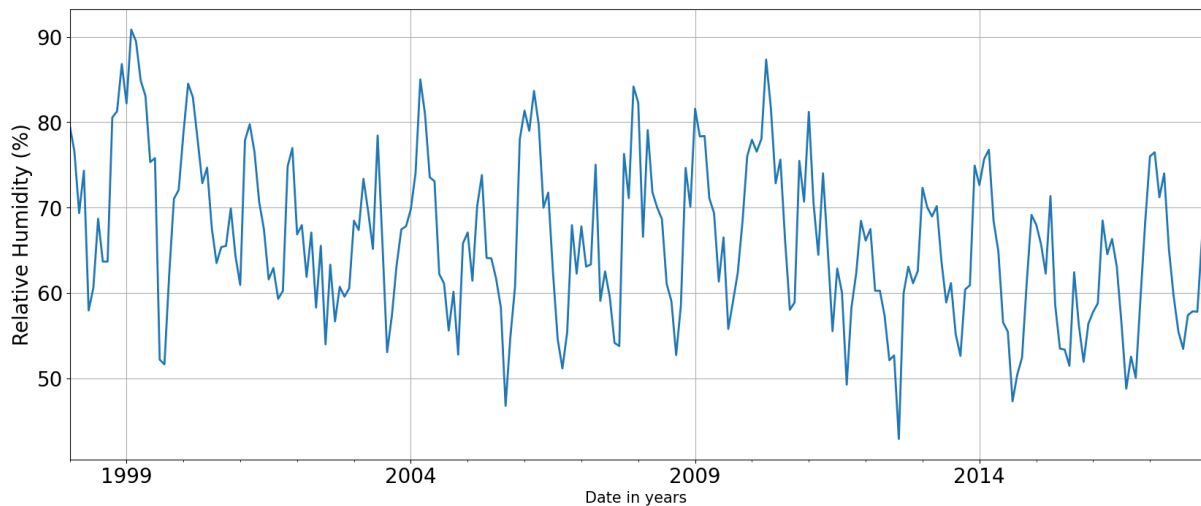


Figure 4.13: Vhembe monthly average relative humidity from year 1998 to 2018.

The data shows that monthly average relative humidity in Vhembe fluctuates significantly over time, with values consistently oscillating between 50% and 90%. Notably, there was a high degree of variability in the early 2000s, when the relative humidity showed both sharp increases and decreases. Post-2009, the fluctuations become somewhat more moderated, although variability remains evident.

Both temperature and relative humidity trends in the Vhembe region show seasonal fluctuations with increased variability from 2004 to 2009. While both temperature and relative humidity are subject to seasonal variations, the temperature trend is more predictable and stable, whereas the relative humidity trend is more erratic and influenced by additional factors beyond simple seasonal changes.

4.3.3 Rainfall

The rainfall data was received from two climate districts. SAWS uses climate districts to collect rainfall. Figure 4.14 shows the locations of climate districts in Vhembe as well as other administrative districts in Limpopo.

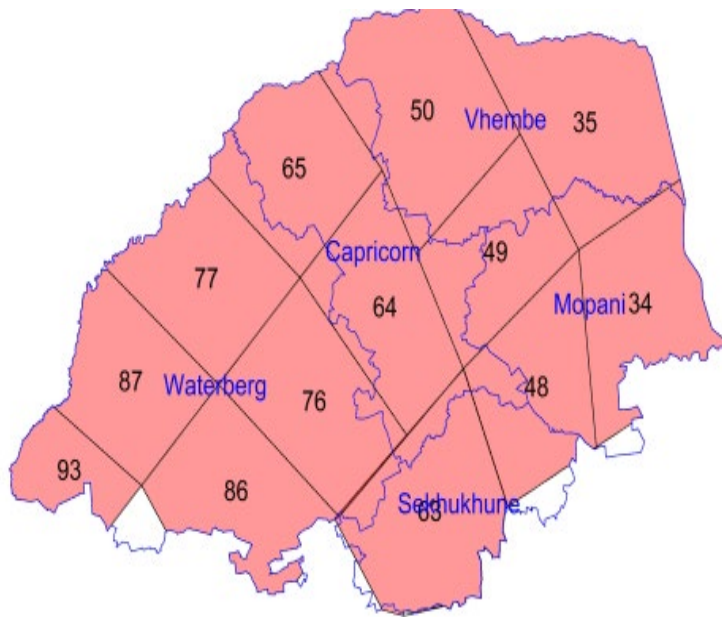


Figure 4.14: Climate and administrative districts of Limpopo.

The climate districts are shown by numbers, while the administrative districts are named. The administrative districts are Capricorn, Greater Sekhukhune, Mopani, Vhembe and Waterberg. The climate districts situated in Vhembe are districts 35 and 50. Figure 4.15 and Figure 4.16 below depict daily rainfall received in Vhembe from districts 35 and 50, respectively.

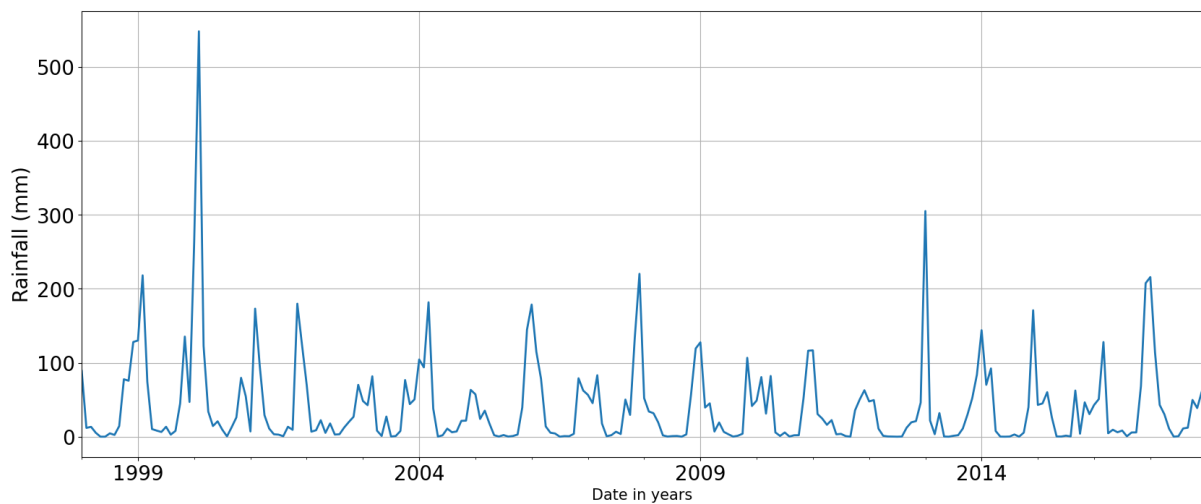


Figure 4.15: daily rainfall from Weather District 35 in Vhembe from year 1998 to 2018.

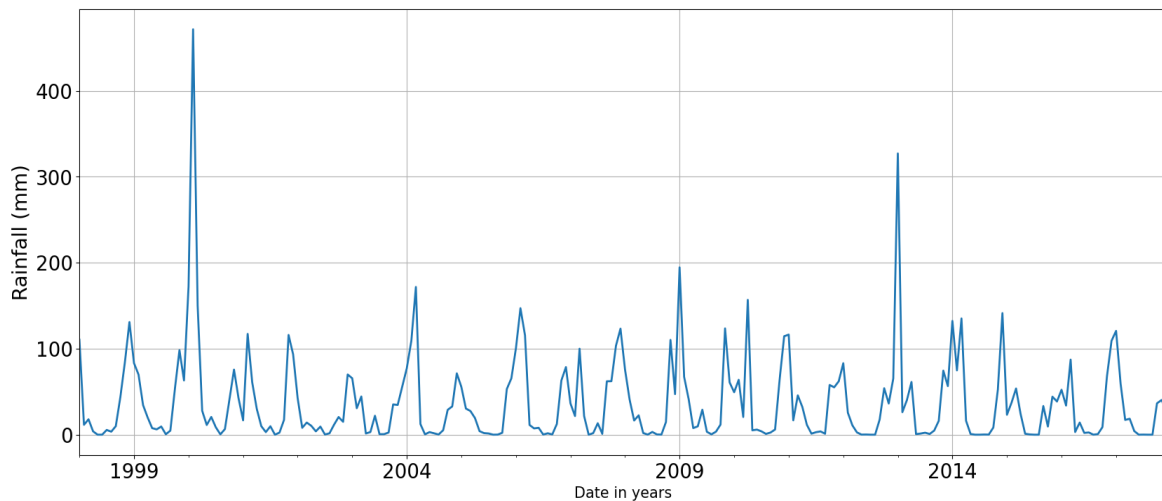


Figure 4.16: daily rainfall from Weather District 50 in Vhembe from year 1998 to 2018.

The rainfall data from both climate districts indicates a pattern of sporadic and intense rainfall events, with pronounced peaks in the years 2000, 2013 and 2016. These peaks represent periods of substantial rainfall, with the most extreme event occurring around the year 2000 when rainfall exceeded 500mm. In the intervals between these peaks, rainfall generally remains low, characterized by smaller, more frequent spikes, suggesting periodic but less intense rainfall events. Generally, it is evident that Vhembe district experiences higher rainfall during the summer months, with little to no rainfall occurring in winter.

4.3.4 Drought Indices

The EDI and AWRI were calculated from daily rainfall records in climate districts 35 and 50 to determine the amount of moisture in the soil. Figures 4.17, 4.18, 4.19 and 4.20 below illustrate EDI and AWRI from climate districts 35 and 50, respectively.

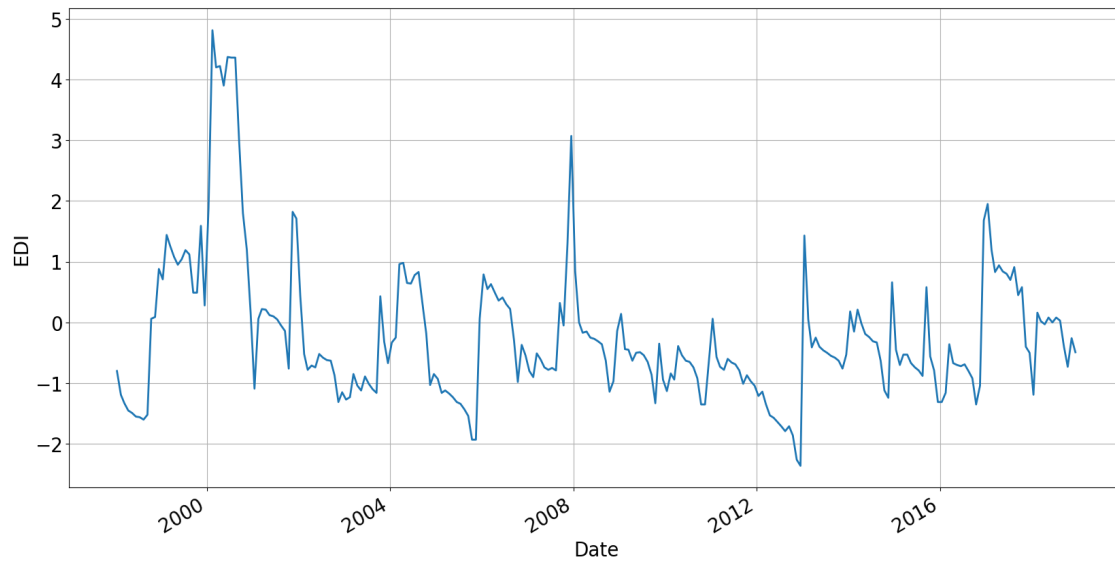


Figure 4.17: EDI from rainfall recorded from year 1998 to 2018 from District 35.

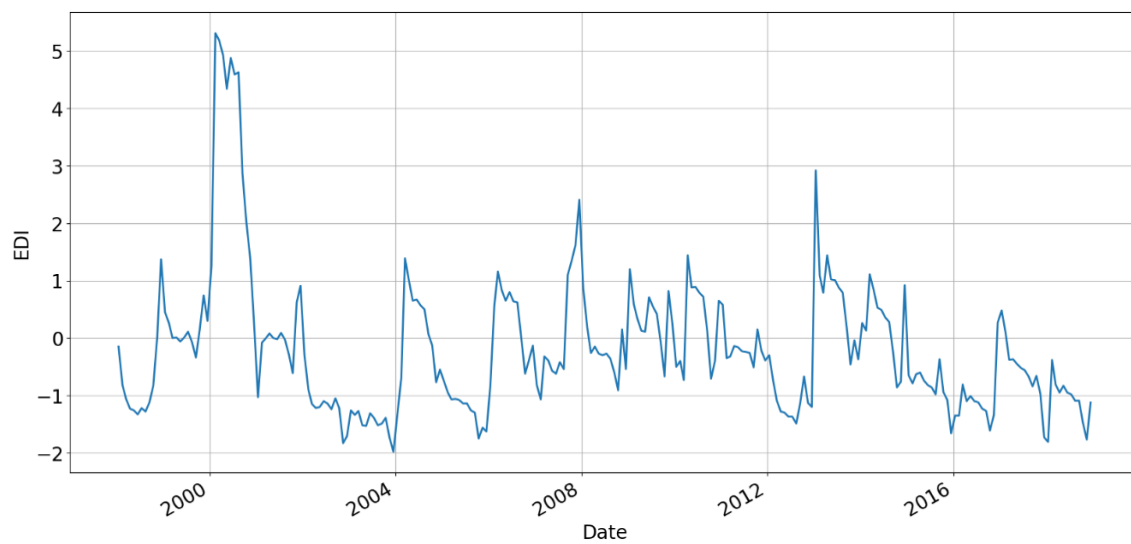


Figure 4.18: EDI from rainfall recorded from year 1998 to 2018 from District 35.

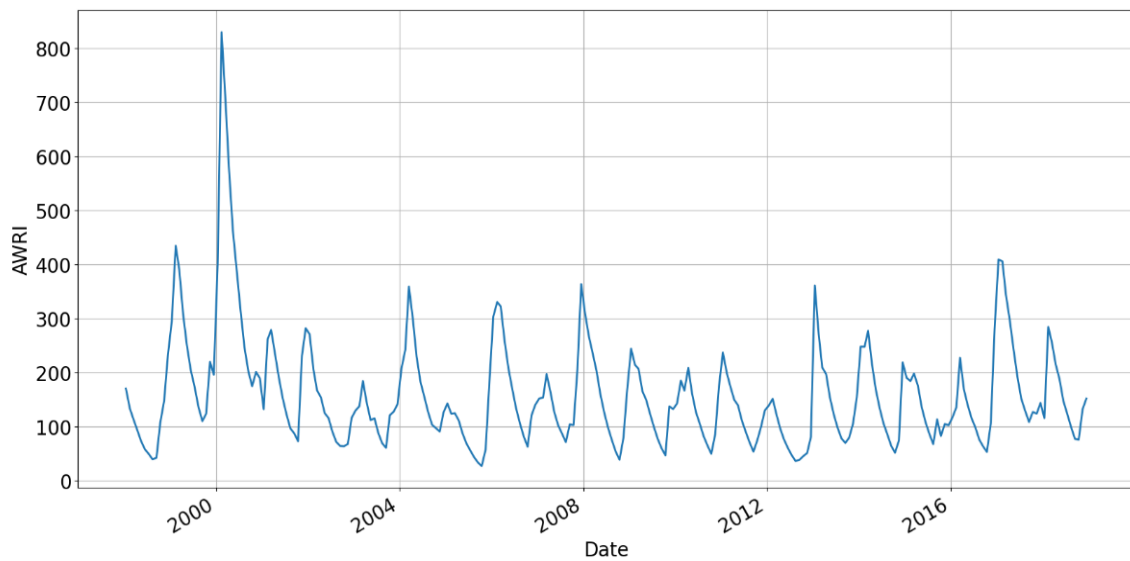


Figure 4.19: AWRI from year 1998 to 2018 from District 35.

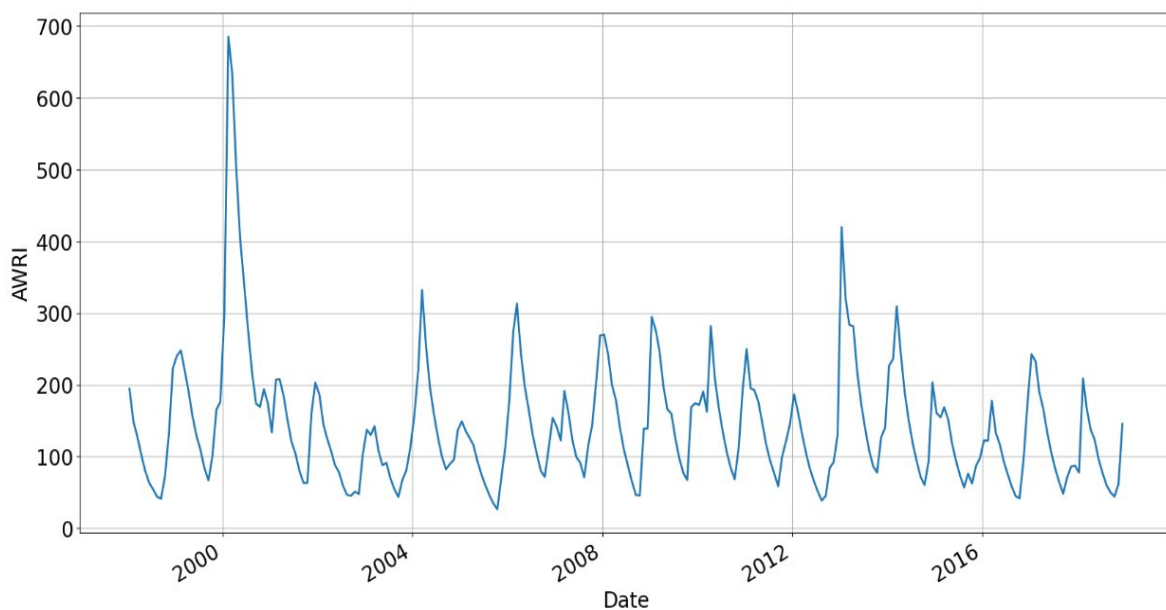


Figure 4.20: AWRI from year 1998 to 2018 from District 50.

From Figures 4.17 to 4.20, it is evident that both districts exhibit similar trends in the AWRI and the EDI. This observation is consistent with the parallel trends in rainfall recorded in both districts. Additionally, it is notable that Vhembe predominantly experiences normal wetness conditions. Drought occurrences in Vhembe are infrequent, with notable exceptions in the years 2003, 2004 and 2006. Conversely, there are several instances of flooding, with the years 2000 and 2001 experiencing extreme flooding, as indicated by EDI values close to 5 in districts 35 and 50, respectively.

4.4 Data Source 3: Structured Health Data

The other secondary data considered in this study is health data. The monthly malaria incidence data was sourced from the Department of Health in Limpopo province of South Africa through SAWS, also from 1998 to 2018 (a 21-year period). Figure 4.21 below reflects Vhembe monthly malaria incidences in comparison to other administrative districts in Limpopo.

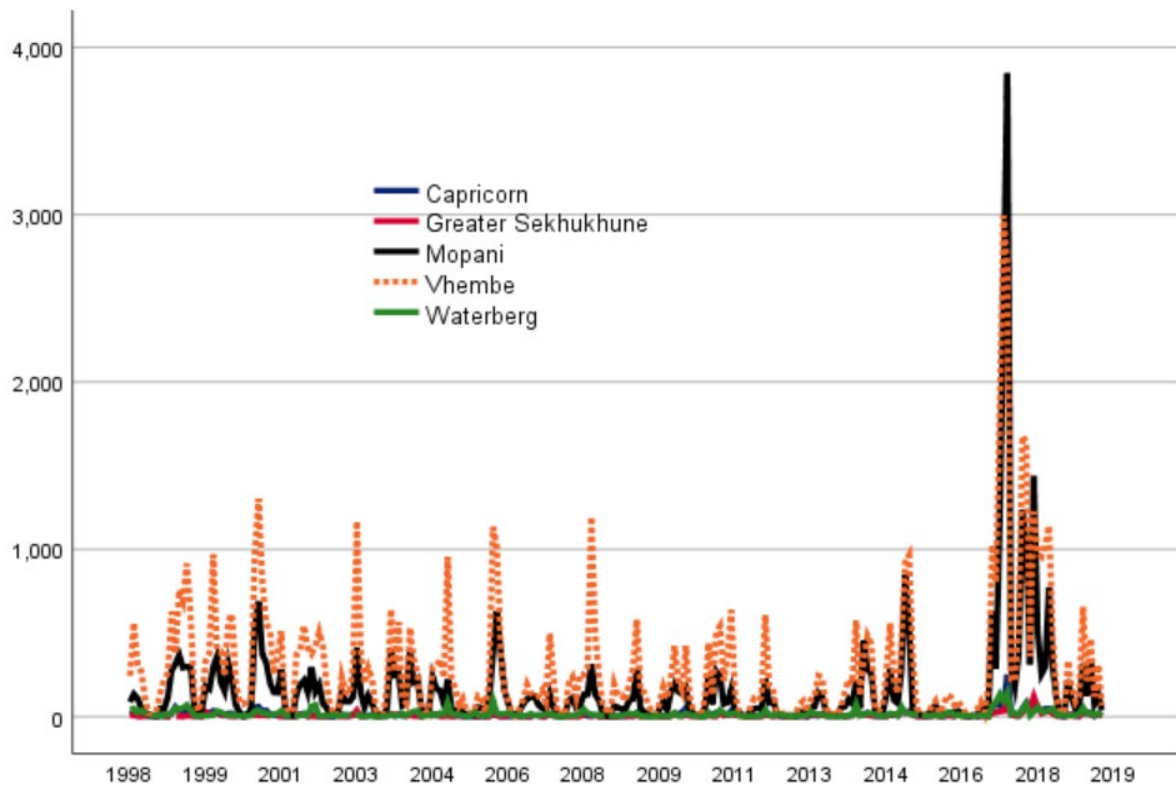


Figure 4.21: Malaria incidences in Vhembe compared to the other four administrative districts in Limpopo from 1998 to 2018.

It can be observed that Vhembe (followed by Mopani) bears the burden of malaria among all the districts in Limpopo province. Of 126503 malaria incidences reported in Limpopo from January 1998 to December 2018, 60% of these incidences were from Vhembe. Vhembe experienced the highest and lowest malaria incidences in 2017 and 2016, respectively.

Figure 4.22 shows malaria incidences in Vhembe according to different months of the year.

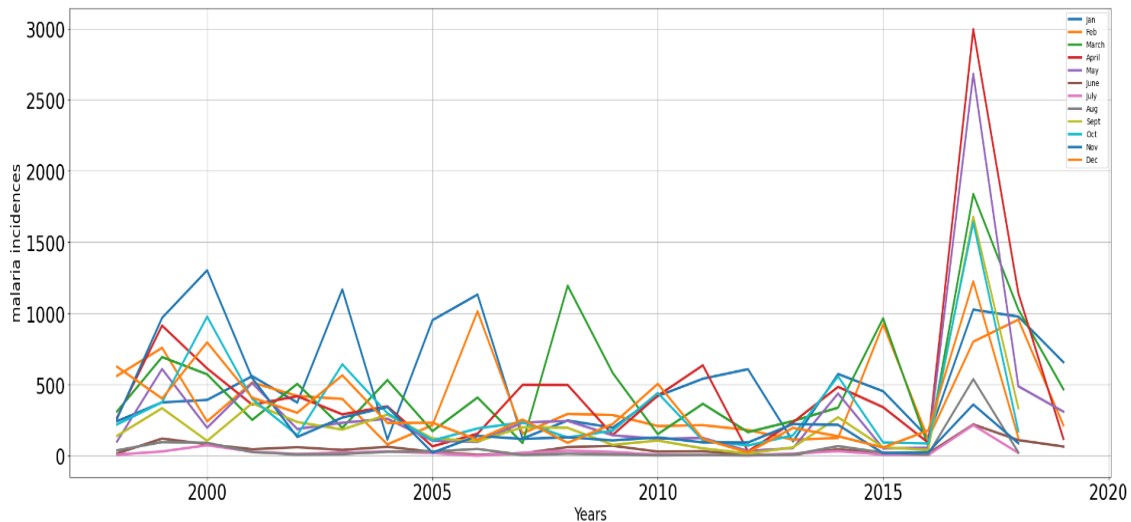


Figure 4.22: malaria incidences in Vhembe according to different months of the year (from 1998 to 2018).

It can be observed that from 1998 to 2018, June, July and August are the months with the least number of malaria incidences. On the other hand, the months that recorded high malaria incidences are January, February, March and April. It can also be noted that Vhembe experiences the highest malaria incidences in summer and autumn (from December to May). This synchronises with the information, depicted in Section 4.2.3.2 (IK indicators for Vhembe participants), given by participants on the knowledge of the seasons regarding malaria outbreaks in Vhembe.

4.5 Conclusion

This chapter provided an overview of the data collection from heterogeneous data sources. While the weather and health data were secondary data sources from the SAWS and the Department of Health, respectively, the IK data was the primary data. The IK data were collected using semi-structured questionnaires,

This chapter revealed the feasibility of developing an EWS for malaria in Limpopo. Participants in the study typically demonstrate extensive residency and a great understanding of their local environment. This prolonged engagement fosters the accumulation of rich IK, leveraged not solely for malaria outbreak prediction but also in farming practices and weather prediction. Moreover, the notable prevalence of smartphone ownership among participants is poised to serve as a pivotal component for the dissemination of malaria alerts, thereby fostering community engagement and response mechanisms.

The chapter analysed the collected IK indicators used by the Vhembe participants to predict malaria outbreaks, along with assessing the community's confidence in using IK. A notable finding was the association between the migration history and the participants' confidence in utilising IK for general purposes and specific purposes such as disease prediction, weather forecasting and farming. Individuals with a longer history in the region displayed greater confidence in the utility of IK in disease prediction, agricultural practices and general applications. Furthermore, the IK data revealed a correlation between the participant's age and their confidence in using IK to forecast disease outbreaks. The older participants showed more confidence in the use of IK for disease prediction. Collecting IK data, historical malaria incidence and weather parameters forms a crucial element of the system's framework, specifically in the domain of malaria outbreak risk knowledge.

The weather data and the AWRI and EDI hydrological drought indices for Vhembe from 1998 to 2018 provided by SAWS were presented. The data include temperature, relative humidity, precipitation and drought indices. Vhembe's monthly average temperatures range from 14.01°C to 26.7°C, showing clear seasonal variations. Rainfall data from the region indicates peak rainfall in 2000 and 2013, with higher precipitation in summer and little to none in winter. The drought indices suggest that Vhembe typically experiences normal wetness, with occasional flooding and some droughts.

Gathering insights into malaria outbreak risk knowledge within the Vhembe community encourages the exploration of a malaria EWS amalgamating IK with contemporary scientific methodologies. The next chapters (Chapters 5 and 6) will explore the prediction of malaria occurrences in Vhembe, building on the data discussed in this chapter.

5 Chapter Five: Predicting Malaria Outcome Using Indigenous Knowledge

5.1 Introduction

The preceding chapter addressed the first component of the Malaria Outbreak Early Warning System (MOEWS), which focused on understanding malaria outbreak risk. In this chapter, along with Chapter 6, the focus shifts to the second component of MOEWS, which is centred on predicting malaria outbreaks in the Vhembe district of Limpopo province, South Africa. The predictions are done utilising two methodologies: the modern scientific approach and the indigenous approach, with this chapter implementing the latter. It details the development of a malaria prediction system that integrates indigenous knowledge (IK) and fuzzy cognitive maps (FCM).

The chapter commences with Section 5.2 which gives an overview of the procedure followed to predict malaria using FCM. Section 5.3 follows to discuss the process of identifying and analysing concepts used in the FCM. Section 5.4 outlines the presentation of concepts in the FCM while Section 5.5 shows the design and building of the FCM. Sections 5.6 and 5.7 detail the inference process and interpretation of the FCM output, respectively. Finally, Section 5.8 provides a conclusion to the chapter.

5.2 Fuzzy Cognitive Map Methodology

The preliminary task was to comprehend the IK domain through a statistical analysis of responses related to IK used in predicting malaria outbreaks. After understanding the domain, the subsequent step was formulating the inputs, processing mechanisms and outputs that constitute the mathematical logic underlying the FCM-based malaria outbreak prediction model. Figure 5.1 displays the process this research study followed to predict malaria outbreaks using FCM.

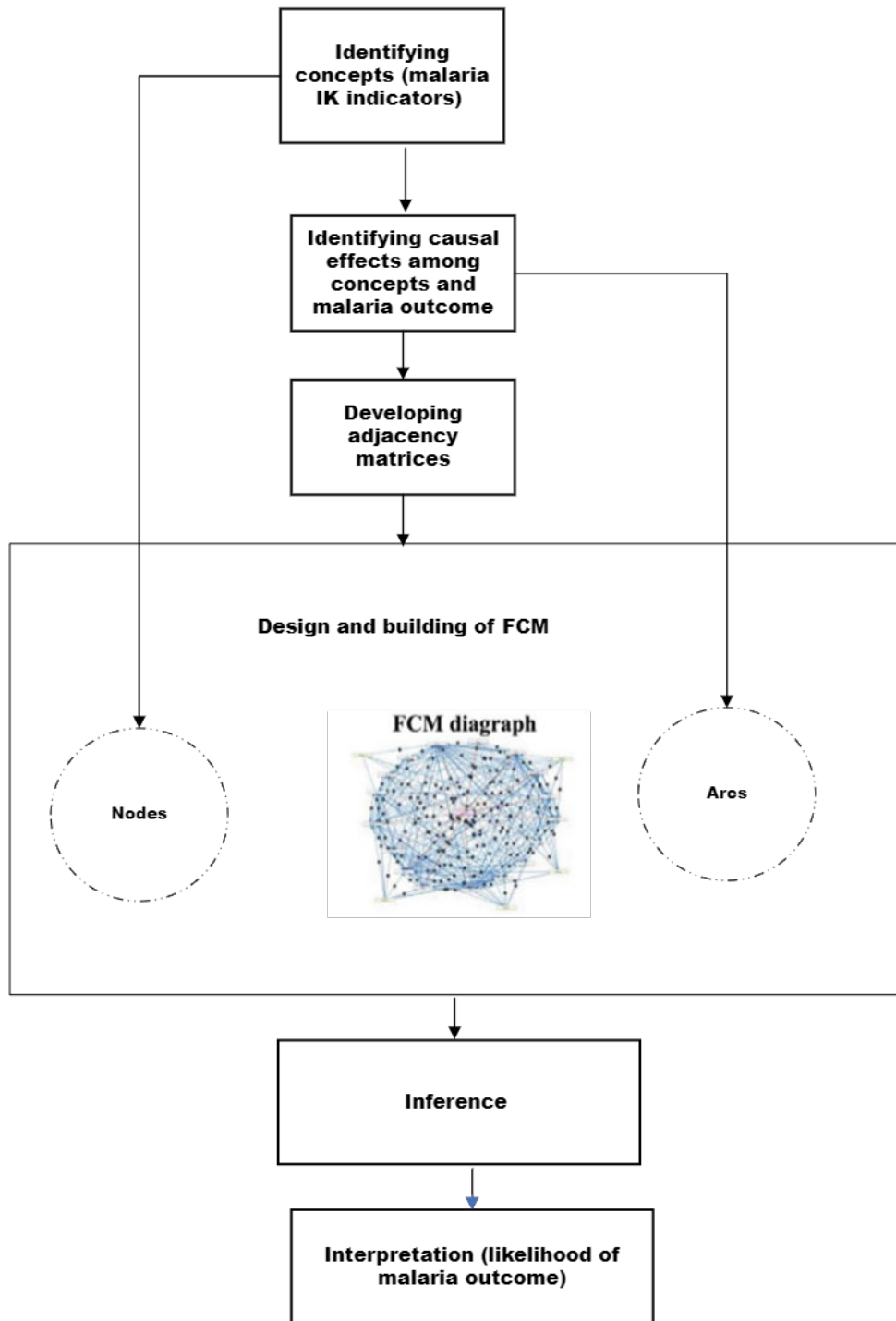


Figure 5.1: Process of predicting malaria outbreaks using FCM.

5.3 Identification and Analysis of Concepts for Malaria Prediction

FCM was used to collect, verify and validate indigenous knowledge indicators (IKIs) for malaria prediction in the Vhembe district in Limpopo province of South Africa. To understand the IK domain for malaria prediction, the first step was identifying the concepts to be used in a fuzzy cognitive map. The IKIs for malaria prediction collected during the data collection stage, documented in Chapter 4, were restructured to form

the FCM concepts. Each IK indicator represents a concept in the FCM malaria prediction model. The list of concepts is shown in Table 5.1 below.

Table 5.1: Malaria outbreak prediction concepts.

	Concepts			
	Summer	Autumn	Winter	Spring
Astronomical and meteorological	C1: Summer floods C2: Summer heavy rainfalls C3: Tilted moon C4: Moon surrounded by clouds. C5: Sun surrounded by the clouds. C6: Clustered stars along the horizon. C7: Moon surrounded by a ring of stars	C1: Autumn lots of rainfalls		C1: Spring rains
Behaviours of Birds	C8: Many birds C9: Swallows		C1: Sight of unusual birds (less/no malaria)	C2: Sight of many birds
Behaviours of Trees/Plants	C10: Black spots on mangos C11: many mangoes/jackals/marula falling from the tree C12: Tree leaves turning yellow (no malaria)		C2: Fig trees 'Muhuyu' not shedding leaves (no/less malaria)	C3: Heavy mangos flowering. C4: Quinine Tree ("Munadzi") dripping C5: "Mofafa grass" has many ticks.
Insect behaviour	C13: Ants gathering food. C14: Sight of butterflies, crows/millipedes	C2: Ants gathering food/storing them underground. C3: Cattle frequently running with tails raised.	C3: Sight of many insects/locusts (no/less malaria)	C6: Ants gathering food.
Animal Behaviour	C15: Sight of many sardines in dams C16: Croaking frogs C17: Cattle running with tails raised up			C7: Many cattle Egrets around cattle.
Myth, Religious Beliefs and others	C18: Severe malaria outbreak after a period of 5 – 7 years. C19: Dirty water in containers/small pools	C4: Dirty water in tins/small pools		
Knowledge of Seasons	C20: Summer high temperature	C5: Autumn high temperature		

As illustrated in Table 5.1, there are 33 concepts in total: 19, 4, 3 and 7 for summer, autumn, winter and spring, respectively. The concepts are classified into Astronomical and Meteorological, Behaviours of Birds, Behaviours of Trees/Plants, Insect behaviour, Myth, Religious Beliefs and others and knowledge of the seasons categories as seen in Chapters 3 and 4.

5.3.1 Determining the Causal Effects Among Concepts

To determine the interrelations among the identified concepts, respondents recognised as IK experts stated their perceptions on the degree of causal effects of IKIs on malaria outcomes, categorised as strong positive, positive, none, negative, or strong negative across four distinct seasons: summer, autumn, winter and spring.

The participants' responses on causal effects of the concepts on malaria outcomes were quantified as percentages. Table 5.2 through to Table 5.5 show statistical representations of IK expert responses for each interacting concept, giving the percentage distribution of responses across categories of causal effect (strong positive, positive, none, negative and strong negative). Mode and mean values derived were used to aggregate perceptions for each concept.

Table 5.2: Causal effects of concepts (IKIs) on malaria outbreak during summer.

IK indicators to malaria outbreak	Causal Effects for Summer Concepts						
	Strong positive	Positive	None	Negative	Strong negative	Mode -1 to 1	Mean -1 to 1
Summer floods	64%	36%				1	0.8
Summer Heavy rainfalls	64%	36%				1	0.8
Summer tilted moon		64%	36%			0.5	0.3
Summer moon surrounded by clouds	36%	64%				0.5	0.7
Summer sun surrounded by the clouds	18%	64%	18%			0.5	0.5
Summer clustered stars along the horizon.		64%	36%			0.5	0.3
Summer moon surrounded by a ring of stars		64%	36%			0	0.3
Summer many birds		64%	36%			0.5	0.3
Summer many swallows	9%	73%	18%			0.5	0.5
Black spots on mangos	27	55%	18%			0	0.5
Too many mangoes/jackals/marula falling from the tree	18%	64%	18%			0.5	0.5
Summer yellow tree leaves			36%	64%		-0.5	-0.3
Ants gathering food		45%	55%			0	0.2
The sight of butterflies/crows/millipedes	18%	64%	18%			0.5	0.5
Croaking frogs	9%	64%	27%			0.5	0.4
The sight of many sardines in dams	27%	55%	18%			0.5	0.5
Cattle frequently running with tails raised up		55%	45%			0.5	0.3
Severe malaria outbreak after 5–7 years		45%	55%			0	0.2
Dirty water in containers/small pools	64%	36%				0.5	0.8
Summer high temperature	36%	55%	9%			0.5	0.6

Table 5.3: Causal effects of concepts (IKIs) on malaria incidences in autumn.

IK indicators to malaria outbreak	Causal effects for Autumn Concepts						
	Strong positive	Positive	None (0)	Negative	Strong negative	Mode -1 to 1	Mean -1 to 1
Autumn heavy rains	45%	55%				1	0.7
Ants gathering food		45%	55%			0	0.2
Cattle frequently running with tails raised up		45%	55%			0	0.2
Dirty water in containers/small pools	64%	36%				0.5	0.8
Autumn high temperature	36%	55%	9%			0.5	0.6

Table 5.4: Causal effects of concepts (IKIs) on malaria outbreak in winter.

IK indicators to malaria outbreak	Causal effects for Winter Concepts						
	Strong Positive (1)	Positive (0.5)	None (0)	Negative (-0.5)	Strong negative (-1)	Mode -1 to 1	Mean -1 to 1
The sight of unusual birds				100%		-0.5	-0.5
The fig "Muhuyu" trees not shedding leaves				82%	18%	-0.5	-0.6
The sight of many insects/locusts				73%	27%	-0.5	-0.6

Table 5.5: Causal effects of concepts (IKIs) on malaria outbreak in spring.

IK indicators to malaria outbreak	Causal effects for Spring Concepts						
	Strong Positive (1)	Positive (0.5)	None (0)	Negative (-0.5)	Strong negative (-1)	Mode -1 to 1	Mean -1 to 1
Spring rains	36%	45%	18%			1	0.6
The sight of many birds	9%	55%	36%			0.5	0.3
Heavy mangos flowering		82%	18%			0.5	0.4
Quinine Tree dripping		55%	45%			0.5	0.2
The "Mofafa grass" having many ticks	18%	55%	27%			0.5	0.4
Ants gathering food		27%	73%			0	0.1
Many cattle Egrets around cattle	9%	55%	36%			0.5	0.3

Table 5.6: Summary of mode and mean of concepts (I_ks) causal effects on malaria according to seasons.

Summer			Autumn			Winter			Spring		
Concept	Mode	Mean	Concept	Mode	Mean	Concept	Mode	Mean	Concept	Mode	Mean
Summer floods	1	0.9	Autumn heavy rains	1	0.7	the sight of unusual birds	-0.5	-0.5	Spring rains	1	0.6
Summer Heavy rainfalls	1	0.8	Ants gathering food	0	0.2	fig "Muhuyu" trees not shedding leaves	-0.5	-0.6	Sight of many birds	0.5	0.3
Summer tilted moon	0.5	0.3	cattle running with tails raised up	0	0.2	Sight of many insects/locusts	-0.5	-0.6	Heavy mangos flowering.	0.5	0.4
Summer Moon surrounded by clouds	0.5	0.7	Dirty water in containers/small pools	0.5	0.8				Quinine Tree ("Munadzi") dripping	0.5	0.2
Summer Sun surrounded by the clouds	0.5	0.5	Autumn temperature	0.5	0.6				"Mofafa grass" having many ticks	1	0.4
Summer Clustered stars along the horizon.	0.5	0.3							Ants gathering food	0	0.1
Summer Moon surrounded by a ring of stars	0	0.3							Many cattle Egrets around cattle.	0.5	0.3
Summer many birds	0.5	0.3									
Summer many swallows	0.5	0.5									
Black spots on mangos	0	0.5									
Too many mangoes/jackals/marula falling from the tree	0.5	0.5									
Summer yellow tree leaves	-0.5	-0.3									
Ants gathering food	0	0.2									
Sight of butterflies/crows/millipedes	0.5	0.5									
Croaking frogs	0.5	0.4									
sight of many sardines in dams	0.5	0.5									
Cattle frequently running with tails raised up	0.5	0.3									
Severe malaria outbreak after 5-7 years	0	0.2									
Dirty water in containers/small pools	0.5	0.8									
Summer high temperature	1	0.6									

Causal effects among the concepts (IKIs to each other) were determined similarly as for concepts (IKIs) on malaria outbreaks, and results are shown in Tables 5.7 to 5.9 according to the seasons: Autumn, Winter, Spring.

Table 5.7: Causal effects between IKIs concepts in Autumn.

Concept to Concept	Causal effects in Autumn						
	Strong Positive (1)	Positive (0.5)	None (0)	Negative (-0.5)	Strong negative (-1)	Mode -1 to 1	Mean -1 to 1
Autumn heavy rains to aunts gathering food			100%			0	0
Autumn heavy rains to cattle running with tails raised up			100%			0	0
Autumn heavy rains to dirty water in containers/small pools	64%	36%				1	0.8
Autumn heavy rains to autumn temperature			100%			0	0
Ants gathering food to autumn heavy rainfall		55%	45%			1	0.3
Ants gathering food to cattle running with tails raised up			100%			0	0
Ants gathering food to dirty water in containers/small pools			100%			0	0
Ants gathering food to autumn temperature			100%			0	0
Cattle frequently running with tails raised up to autumn heavy rains		64%	36%			0.5	0.3
Cattle frequently running with tails raised to ants gathering food			100%			0	0
Cattle frequently running with tails raised to dirty water in containers/small pools			100%			0	0
Cattle frequently running with tails raised to autumn temperature			100%			0	0
Dirty water in containers/small pools to autumn heavy rains			100%			0	0
Dirty water in containers/small pools to ants gathering food			100%			0	0
Dirty water in containers/small pools to cattle running with tails raised up			100%			0	0
Dirty water in containers/small pools to autumn temperature			100%			0	0
Autumn temperature to autumn heavy rains			100%			0	0
Autumn temperature to ants gathering food			100%			0	0
Autumn temperature to cattle running with tails raised up			100%			0	0
Autumn temperature to dirty water in containers/small pools			64%	36%		0	-0.2

Table 5.8: Causal effects between IKIs concepts in Winter.

Concept to Concept	Causal effects of concepts in Winter						
	Strong Positive (1)	Positive (0.5)	None (0)	Negative (-0.5)	Strong negative (-1)	Mode -1 to 1	Mean -1 to 1
Sight of unusual birds to fig tree not shedding leaves			100%			0	0
Sight of unusual birds to sight of many insects/locusts			100%			0	0
Fig "Muhuyu" trees not shedding leaves to sight of unusual birds			100%			0	0
Fig "Muhuyu" trees not shedding leaves to sight of many insects/locust			100%			0	0
Sight of many insects/locusts to sight of unusual birds			100%			0	0
Sight of many insects/locusts to fig tree not shedding leave			100%			0	0

Table 5.9: Causal effects between IKIs concepts in Spring.

Concepts to concepts	Causal effects in Spring						
	Strong Positive (1)	Positive (0.5)	None (0)	Negative (-0.5)	Strong negative (-1)	Mode -1 to 1	Mean -1 to 1
Spring rains to sight of many birds			100%			0	0
Spring rains to heavy mangos flowering	9%	64%	27%			0.5	0.4
Spring rains to 'Mofafa' grass having many ticks			100%			0	0
Spring rains to ants gathering food			100%			0	0
Spring rains to many cattle Egrets around cattle			100%			0	0
Spring rains to quinine Tree dripping			100%			0	0
Sight of many birds to spring rains		64%	36%			0.5	0.3
Sight of many birds to heavy mangos flowering			100%			0	0
Sight of many birds to 'Mofafa' grass having many ticks			100%			0	0
Sight of many birds to ants gathering food			100%			0	0
Sight of many birds to many cattle Egrets around cattle			100%			0	0
Sight of many birds to quinine tree dripping			100%			0	0
Heavy mangos flowering to spring rains			100%			0	0
Heavy mangos flowering to sight of many birds			100%			0	0
Heavy mangos flowering to 'Mofafa' grass having many ticks			100%			0	0
Heavy mangos flowering to ants gathering food			100%			0	0
Heavy mangos flowering to many cattle Egrets around cattle			100%			0	0
Heavy mangos flowering to quinine Tree dripping			100%			0	0



'Mofafa' grass having many ticks to spring rains	18%	64%	18%			0.5	0.5
'Mofafa' grass having many ticks to sight of many birds			100%			0	0
'Mofafa' grass having many ticks to ants gathering food			100%			0	0
'Mofafa' grass" having many ticks to many cattle Egrets around cattle			100%			0	0
'Mofafa' grass having many ticks to quinine tree dripping			100%			0	0
"Mofafa grass" having many ticks to heavy mangos flowering			100%			0	0
Ants gathering food to spring rains		36%	64%			0.5	0.2
Ants gathering food to sight of many birds			100%			0	0
Ants gathering food to Mofafa grass" having many ticks			100%			0	0
Ants gathering food to many cattle Egrets around cattle			100%			0	0
Ants gathering food to heavy mangos flowering			100%			0	0
Ants gathering food to quinine Tree dripping			100%			0	0
Many cattle Egrets around cattle to spring rains	9%	55%	36%			0.5	0.3
Many cattle Egrets around cattle to sight of many birds			100%			0	0
Many cattle Egrets around cattle to Mofafa grass" having many ticks			100%			0	0
Many cattle Egrets around cattle to many cattle Egrets around cattle			100%			0	0
Many cattle Egrets around cattle to quinine Tree dripping			100%			0	0
Many cattle Egrets around cattle to ants gathering food			100%			0	0
Quinine Tree dripping to spring rains			100%			0	0
Quinine Tree dripping to sight of many birds			100%			0	0
Quinine Tree dripping to mofafa grass" having many ticks			100%			0	0
Quinine Tree dripping to many cattle Egrets around cattle			100%			0	0
Quinine Tree dripping to ants gathering food			100%			0	0
Quinine Tree dripping to heavy mangos flowering			100%			0	0

The causal effects between the concepts for summer were calculated in the same manner as for the other seasons and were used to build a summer weighted adjacency matrix discussed in the next sections.

The representation of the aggregated responses of participants using the mean instead of the mode depicted a good aggregate of causal effects; hence the mean was preferred as the best statistic to represent the causal effects between the concepts.

5.3.2 Building Weighted Adjacency Matrices

To build weighted adjacency matrices, the researcher assigned numeric values to the relationships between concepts based on their degree of causality. These values, ranging from -1 to 1, represent the strength and direction of influence between concepts, forming a matrix that quantifies these interactions. The subsequent sections elaborate on this process of building weighted adjacency matrices.

5.4 Representation of Concepts in Fuzzy Cognitive Maps

The concepts (for both IKIs and malaria outbreak) identified during the understanding of IKIs for malaria prediction were uniquely structured by their position in an n -state vector $(C_1, C_2, \dots, C_n)$, where n is the number of concepts, the last of which will be an actual malaria outbreak.

Each indicator represents a concept or node in an FCM as mentioned in Section 5.3.1. Each concept is characterized by a fuzzy value, P , restricted to $P \in [0, 1]$. Zero indicates no absence, an increase from zero to 1 indicates an increase in presence while 1 indicates full presence as discussed in Chapter 2.

5.5 Representation of Causal Effects of Concepts Using Weighted Adjacency Matrices

The relationships between the concepts were described using a degree of causality (Apostolopoulos and Groumpos, 2023). The degree of influence/impact/causality, also known as the weight (w), between concepts stated by the participants in Section 5.3.1 using linguistic variables, was converted into values in the interval $[-1, 1]$, (Mbele and Masinde, 2020), categorised as:

Strong positive	$(0.5, 1]$
-----------------	------------

Positive	$(0, 0.5]$
None	0
Negative	$[-0.5, 0)$
Strong negative	$[-1, -0.5)$

The weighted $n \times n$ adjacency matrices consisting of the weights are shown in Tables 5.10 to Table 5.13, representing the relations between IK concepts with each other and also with malaria outbreaks, season-by-season.

Table 5.10: Weighted adjacency matrix for Summer.

Concepts	Symbol	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15	C16	C17	C18	C19	C20	C21
Summer floods	C1	0	0	0	0	0	0	0	0.0	0	0	0	0	0	0	0	0	0	0	0.7	0	0.8
Summer heavy rainfalls	C2	0.8	0	0	0	0	0	0	0.0	0	0	0	0	0	0	0	0	0	0	0.7	0	0.8
Summer tilted moon	C3	0	0.6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.3
Summer moon surrounded by clouds	C4	0	0.8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.7
Summer sun surrounded by clouds	C5	0	0.8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.5
Summer clustered stars along the horizon	C6	0	0.6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.3
Summer moon surrounded by a ring of stars	C7	0	0.4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.3
Summer many birds	C8	0	0.4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.3
Summer many swallows	C9	0	0.5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.4
Black spots on mangos	C10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.5
Too many mangoes / jackals / marula falling from the tree	C11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.5
Summer yellow tree leaves	C12	0	-0.1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-0.3
Ants gathering food	C13	0	0.3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.2
Sight of butterflies/crows/millipedes	C14	0	0.4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.4
Croaking frogs	C15	0	0.4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.3
Sight of many sardines in dams	C16	0	0.5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.4
Cattle frequently running with tails raised up	C17	0	0.3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.3
Severe malaria outbreak after 5-7 years	C18	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.1
Dirty water in containers/small pools	C19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.8
Summer high temperature	C20	0	0	0	0	0	0	0	0	0	0	0	0.1	0	0	0	0	0	0	-0.2	0	0.7
Malaria outbreak	C21	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Table 5.11: Weighted adjacency matrix for Autumn.

Concepts	Symbol	C1	C2	C3	C4	C5	C6
Autumn heavy rains	C1	0	0	0	0.8	0	0.7
Ants gathering food	C2	0.3	0	0	0	0	0.2
cattle running with tails raised up	C3	0.3	0	0	0	0	0.2
Dirty water in containers/small pools	C4	0.8	0	0	0	0	0.8
Autumn temperature	C5	0	0	0	-0.2	0	0.6
Malaria outbreak	C6	0	0	0	0	0	0

Table 5.12: Weighted adjacency matrix for Winter.

Concepts	Symbol	C1	C2	C3	C4
sight of unusual birds	C1	0	0	0	-0.5
fig "Muhuyu" trees not shedding leaves	C2	0	0	0	-0.6
Sight of many insects/locusts	C3	0	0	0	-0.6
Malaria outbreak	C4	0	0	0	0

Table 5.13: Weighted adjacency matrix for Spring

Concepts	Symbol	C1	C2	C3	C4	C5	C6	C7	C8
Spring rains	C1	0	0	0.3	0	0	0	0	0.6
Sight of many birds	C2	0.3	0	0	0	0	0	0	0.3
Heavy mangos flowering	C3	0	0	0	0	0	0	0	0.4
Quinine tree ("Munadzi") dripping	C4	0	0	0	0	0	0	0	0.2
"Mofafa grass" having many ticks	C5	0.5	0	0	0	0	0	0	0.4
Ants gathering food	C6	0.1	0	0	0	0	0	0	0.2
Many cattle Egrets around cattle	C7	0.3	0	0	0	0	0	0	0.3
Malaria outbreak	C8	0	0	0	0	0	0	0	0

The causal effect of any concept C_i on another concept C_j is denoted by w_{ij} ($1 \leq i \leq n$; $1 \leq j \leq n$) in the adjacency matrix W :

$$W = \begin{pmatrix} w_{11} & \dots & w_{1j} \dots \\ \vdots & \dots & \vdots \\ w_{i1} & \dots & w_{ij} \dots \end{pmatrix} \quad \text{Equation 5.1}$$

In the weighted adjacency matrices, the names of all the concepts were written as row headings and concept symbols, $C_1, C_2, \dots, C_n$, were written as column headings where each cell contains a value indicating the weight (w_{ij}). The value of w_{ij} indicates how strongly concept C_i (row i) influences concept C_j (column j). The sign of w_{ij} indicates whether the relationship between concepts C_i and C_j is direct or inverse (Hoyos, Aguilar and Toro, 2022). A positive weight indicates a positive causality

between concepts, whereas a negative weight indicates a negative causality (Napoles *et al.*, 2028). For example, from the weighted adjacency matrix representing summer knowledge (Table 5.10), $w_{1,21}$ has a positive value (0.8), meaning that an increase in floods strongly predicts an increase in malaria outbreaks. On the other hand, $w_{12,21}$ is a negative value (-0.3) indicating an inverse relationship; an increase in the yellow tree leaves (C12) results in a decrease in malaria outbreaks (C21). Yellow tree leaves in summer mean drought, hence fewer malaria outbreaks. The diagonal of the matrix is zero since it is assumed that the concept does not have a causal effect on itself, and thus, $w_{ii} = 0$. The concepts with no relationship were also given the value of zero.

5.6 Design and Building of Fuzzy Cognitive Maps

Utilizing the weighted adjacency matrices presented in Tables 5.10 to 5.13, fuzzy cognitive maps were constructed using FCM Expert software to represent the relationships among concepts visually (Nápoles *et al.*, 2018). The nodes of the FCM represented the concepts, while the weights/causal effects correspond to the values on the arcs of the FCM. Figure 5.2 below shows a FCM for Autumn.

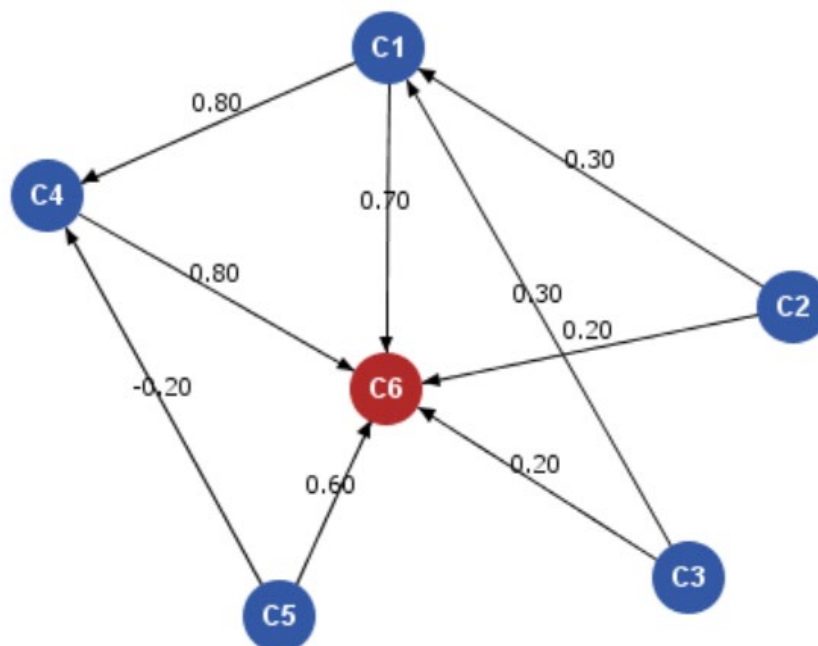


Figure 5.2: FCM for autumn (Key: C1=autumn heavy rainfalls, C2= ants gathering food, C3=cattle running with their tails raised, C4=dirty water in the containers/small pools, C5=autumn temperature, C6=malaria outbreak).

The FCM indices such as the significance of in-degree (sum of weights of the incoming links to a concept), out-degree (sum of weights of outgoing links from a concept) and centrality of the concepts (sum of weights of the outgoing and ingoing links) (Nozari *et al.*, 2021) were calculated using the mental modeler software. Table 5.14 demonstrates the FCM indices for Autumn concepts.

Table 5.14: Analysis of the importance of nodes/concepts in the FCM for Autumn.

Concepts	Symbol	Concept type	In-degree	Out-degree	Centrality
Autumn heavy rainfalls	C1	ordinary	0.6	1.5	2.1
Ants gathering food	C2	driver	0	0.5	0.5
Cattle running with their tails raised up	C3	driver	0	0.5	0.5
Dirty water in the containers/small pools	C4	ordinary	0.8	0.8	1.6
Autumn temperature	C5	driver	0	0.4	0.4
Malaria outbreak	C6	receiver	1.9	0	1.9

The significance of in-degree, out-degree and centrality of the winter concepts is illustrated in the table below. High centrality for a concept indicates that the weights and connections are higher, thus making the concept more central to the system. Concept centrality emphasises the importance of the concepts in the FCM (Gray and Cox, 2022). For the autumn concepts, the concepts with high centrality are Autumn heavy rainfalls and Dirty water in the containers/small pools.

Figure 5.3 demonstrates the FCM for Summer.

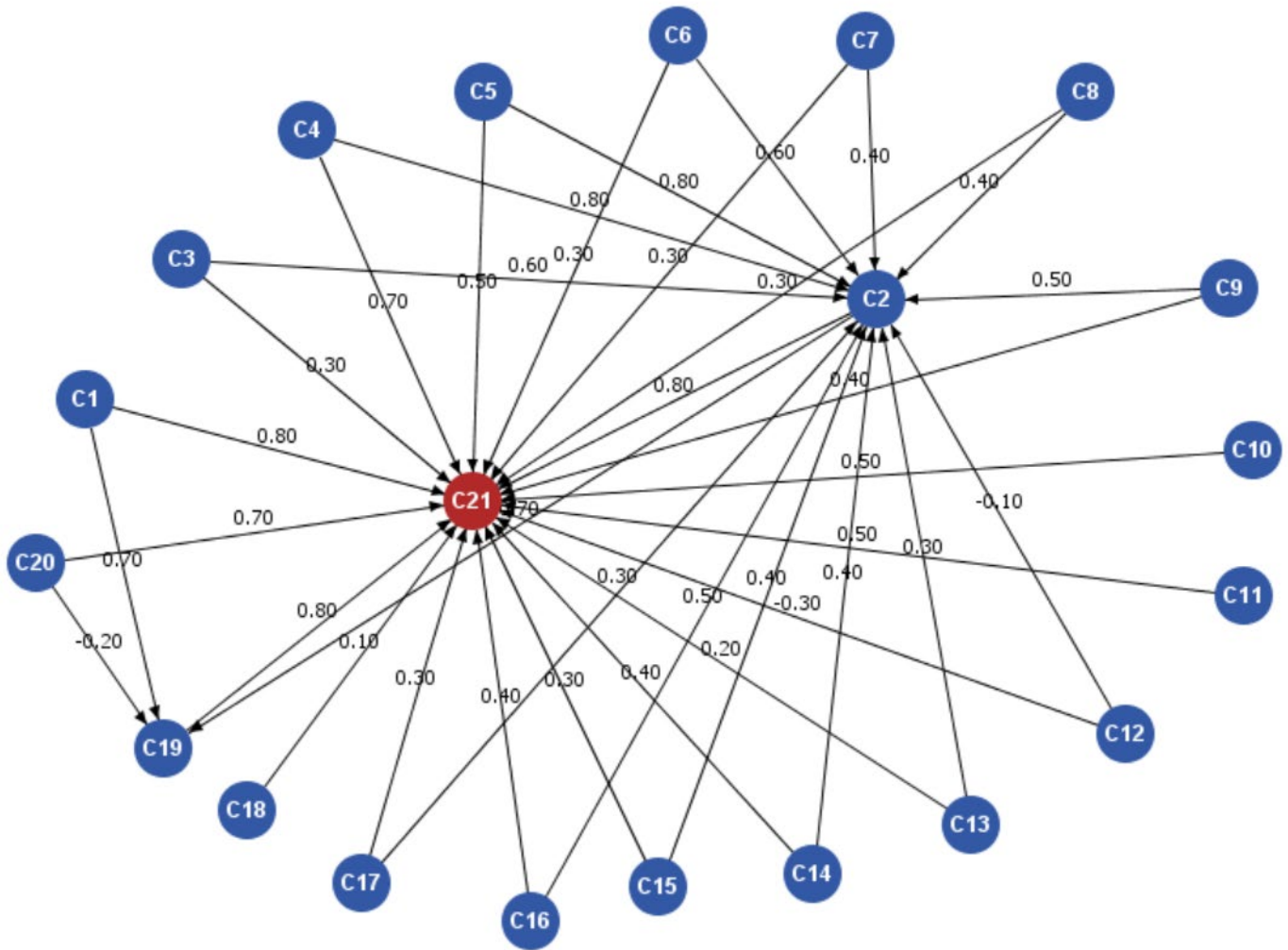


Figure 5.3: FCM for Summer.

The significance of the out-degree, in-degree, and centrality of concepts for the summer FCM model are shown in Table 5.15.

Table 5.15: Analysis of the importance of nodes/concepts in summer FCM.

Concept	Symbol	Concept type	In-degree	Out-degree	Centrality
Summer floods	C1	driver	0	1.6	1.6
Summer heavy rainfalls	C2	ordinary	6.1	1.5	7.61
Summer tilted moon	C3	driver	0	0.899	0.899
Summer moon surrounded by clouds	C4	driver	0	1.5	1.5
Summer sun surrounded by clouds	C5	driver	0	1.3	1.3
Summer clustered stars along the horizon	C6	driver	0	0.899	0.899
Summer moon surrounded by a ring of stars	C7	driver	0	0.7	0.7
Summer many birds	C8	ordinary	0.1	0.7	0.799
Summer many swallows	C9	driver	0	0.9	0.9

Black spots on mangos	C10	driver	0	0.5	0.5
Too many mangoes / jackals / marula falling from the tree	C11	driver	0	0.5	0.5
Summer yellow tree leaves	C12	driver	0.1	0.5	0.5
Ants gathering food	C13	driver	0	0.6	0.6
Sight of butterflies / crows / millipedes	C14	driver	0	0.81	0.81
Croaking frogs	C15	driver	0	0.7	0.7
Sight of many sardines in dams	C16	driver	0	0.9	0.9
Cattle frequently running with tails raised up	C17	driver	0	0.6	0.6
Severe malaria outbreak after 5-7 years	C18	driver	0	0.1	0.1
Dirty water in containers/small pools	C19	ordinary	1.599	0.8	2.4
Summer high temperature	C20	driver	0	1	1
Malaria outbreak	C21	receiver	9	0	9

For summer, rainfalls, floods, dirty water in the containers or small pools, the moon and sun surrounded by clouds and summer temperature are crucial IKIs for predicting malaria in Vhembe.

A similar process was followed to design and build FCMs and FCM indices for Winter and Spring. Figures 5.4 and 5.5 illustrate the FCM for Winter and Spring, respectively.

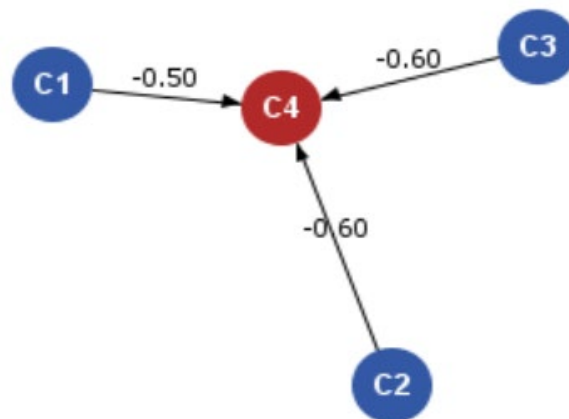


Figure 5.4: Winter FCM (Key: C1= Sight of unusual birds, C2= Fig “Muhuyu” trees not shedding leaves, C3= Sight of many insects/locusts, C4 =Malaria outbreak).

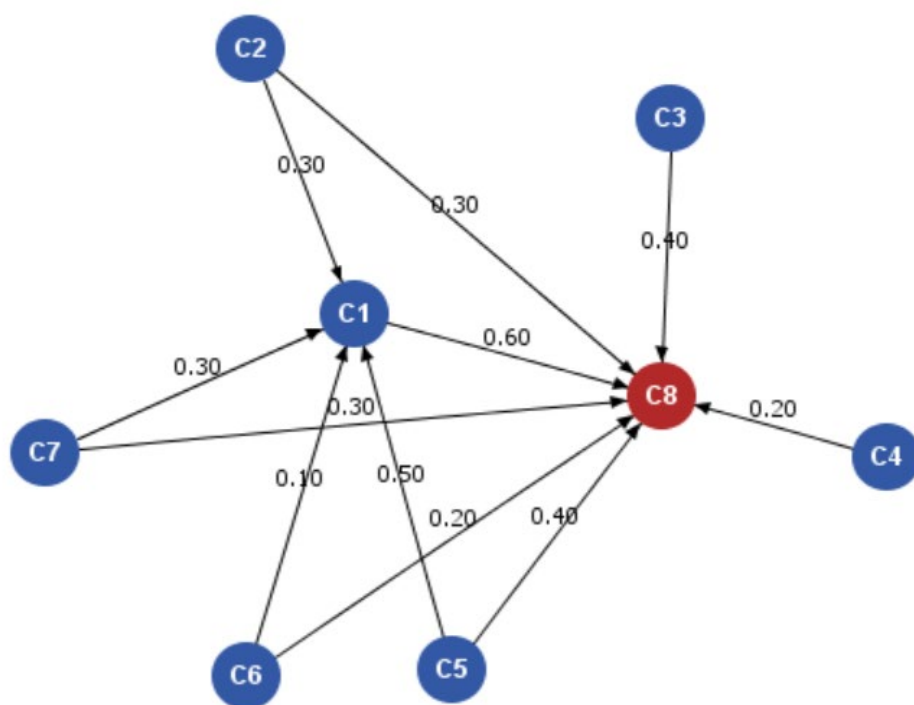


Figure 5.5: FCM for Spring (Key: C1= Spring rains, C2= Sight of many birds, C3= Heavy mangos flowering, C4= Quinine tree (“Munadzi”) dripping, C5= “Mofafa grass” having many ticks, C6= Ants gathering food, C7 = Many cattle Egrets around cattle, C8 =Malaria outbreak.

Table 5.16 shows the FCM indices for Winter, while Table 5.17 depicts the FCM indices for Spring.

Table 5.16: Analysis of the importance of nodes/concepts in Winter FCM.

Concepts	Symbol	Concept type	In-degree	Out-degree	Centrality
Sight of unusual birds	C1	Driver	0	0.5	0.5
Fig “Muhuyu” tress not shedding leaves	C2	Driver	0	0.6	0.6
Sight of many insects/locusts	C3	Driver	0	0.6	0.6
Malaria outbreak	C4	Receiver	1.7	0	1.7

Table 5.17: Analysis of the importance of nodes/concepts in Spring FCM.

Concepts	Symbol	Concept type	In-degree	Out-degree	Centrality
Spring rains	C1	ordinary	1.2	0.6	1.8
Sight of many birds	C2	driver	0	0.6	0.6
Heavy mangos flowering	C3	driver	0	0.4	0.4
Quinine tree (“Munadzi”) dripping	C4	driver	0	0.2	0.2

"Mofafa grass" having many ticks	C5	driver	0	1	1
Ants gathering food	C6	driver	0	0.3	0.3
Many cattle Egrets around cattle	C7	driver	0	0.6	0.6
Malaria outbreak	C8	receiver	2.5	0	2.5

The main predictors of malaria in winter were 'Fig "Muhuyu" trees not shedding leaves' and 'Sight of insects/locusts' while spring rainfalls and 'mofafa grass having many ticks' were found important for malaria prediction in spring.

5.7 Inference

The prediction of malaria outbreaks was conducted by running simulations of scenarios through an inference process. The inference process is a decision-making method that applies fuzzy logic to convert given inputs into corresponding outputs (Felix *et al.*, 2019). This was done using the knowledge represented in the FCMs generated in previous sections.

5.7.1 Representation of the input concepts vector

The input to the FCM model for malaria prediction comprises a list of values representing observed malaria IKIs. IK experts were tasked with reporting IKIs observed in autumn 2023, which were captured using a mobile app. The values of the concepts were determined using the conversion below:

None (not observed)	0 (Default)
Very low	(0, 0.3]
Low	(0.3, 0.5]
Moderate	(0.5, 0.75]
High	(0.75, 1]

The number of concepts determines the size of the input concept vector and the number of rows or columns of the corresponding weighted adjacency matrix. Subsequently, the values of concepts are straightforwardly retrieved from the indices of the values in the input concept vector. This process of determining a number of concepts and concept values is shown in Figure 5.6 below.

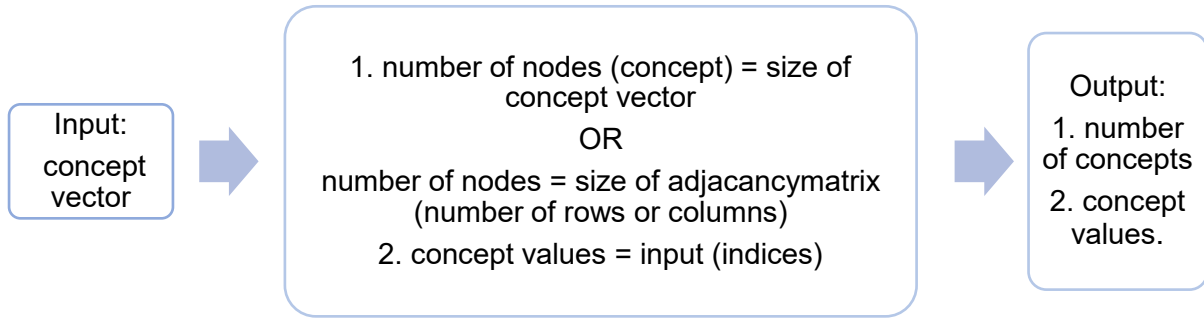


Figure 5.6: Process of determining number of concepts and concepts values in FCM (Mwagha, 2016).

5.7.2 Representation of Causal Relationships between Concepts

The causal effects among the concepts (weights) are represented in the adjacency matrices. The concepts have different causal effects on malaria outbreaks depending on the season, therefore, there are four weighted adjacency matrices distinguished by the four weather seasons (summer, autumn, winter, and spring). During the simulations, the weights from a relevant season are considered.

5.7.3 Activation of New Concepts States and States of Equilibrium

Concepts are activated using the FCM for the corresponding season, and the simulation is run. The value of each concept, $C(P)$, (which indicates concept's level of presence) is influenced by the values of the other connected concepts with their corresponding causal weights (Nápoles *et al.*, 2018). An initial activation state of a concept is denoted with C_0 (Papageorgiou *et al.* 2018). An activation formula is used to compute new values of all concepts for every step of the FCM, successively multiplying the state vector by the weighted adjacency matrix (Papageorgiou *et al.*, 2009), and adding this to the values of the concepts in the previous iteration, as depicted in Equation 5.2 below:

$$C_i^{(t+1)} = f(C_i^t) + \sum_{j=1}^N C_j^t W_{ij} \quad \text{Equation 5.2}$$

where C_i^t is the value of concept C_i at step t , C_j^t is the value of concept C_j at step t and W_{ij} is the weight of the interconnection from concept C_i to concept C_j . The function f is a sigmoid-shaped function which acts as a threshold function to clamp the results in the interval $[0,1]$. The sigmoid function was chosen for its simplicity and popular use in modelling concepts in disease prediction (Groumpos, 2021; Hoyos, Aguilar and Toro, 2022). Unlike other activation functions, such as the hyperbolic tangent which

squashes values within a range of $[-1, 1]$, the sigmoid function restricts values to the range of $[0, 1]$ (Felix *et al.*, 2019). This is particularly useful in this research study, as the modelled concepts range from 0, indicating the IKI is not present or observed, to 1, indicating the full presence of the observed IKI. In each of the test scenarios, there is an initial vector V_i , representing the values of the observed IKI at a given time of the process. If a certain IKI is not observed, it takes a value of zero in the initial vector, or it takes a value between 0 and 1 as explained previously. The final vector V_f , represents the last state achieved through the inference procedure. The inference procedure calculates new concept values, and it continues until the system reaches a steady state equilibrium (Felix *et al.*, 2019). In this research study, the fixed-point attractor of 0.001 was used to identify the steady state. When the difference between the values of the concepts in the current and previous iteration is 0.001 or less, it is taken as an indication that the values of the concepts will not evolve anymore and that the system has converged. The steady state in this case can be interpreted as a state in which the IKIs do not evolve further.

For a scenario run, below is an input vector, denoted by V_i , for IKIs reported by IK experts in the autumn season in 2023 in the Vhembe district. Only autumn heavy rainfall, dirty water in the containers/small pools and autumn temperature (C1, C4, C5) were observed.

$$V_i = \{0.5, 0.0, 0.0, 0.2, 0.3, 0.0\}$$

Table 5.18 and Figure 5.7 below show the initial values for concepts and the final computed concept values.

Table 5.18: Concept values from the initial to the final iteration of the inference process.

Step	C1	C2	C3	C4	C5	C5
0	0,5	0,0	0,0	0,2	0,3	0,0
1	0,6225	0,5	0,5	0,6318	0,5744	0,5854
2	0,7155	0,6225	0,6225	0,734	0,6398	0,738
3	0,7482	0,6508	0,6508	0,7647	0,6547	0,7737
4	0,7598	0,6572	0,6572	0,7742	0,6581	0,7826
5	0,7598	0,6586	0,6586	0,777	0,6588	0,785
6	0,7604	0,659	0,659	0,7778	0,659	0,7856

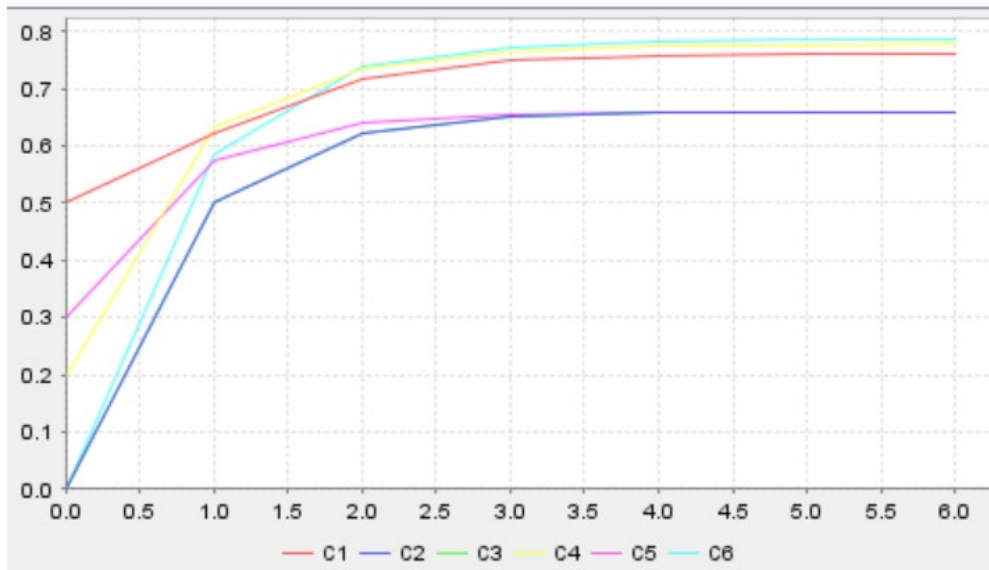


Figure 5.7: Graphical illustration showing initial concept values on the y-axis and changes in concept values on the x-axis across iterations of the inference process.

5.8 Interpretation: The Likelihood of Malaria Outcomes

The final vector Vf of the inference process represents the last state of the concept vector which includes the predicted value of the output concept, a malaria outbreak, also known as the decision concept. This value is in the closet set of $[0,1]$, representing the level of the expected malaria outbreak. To interpret the results, the value of the malaria concept was formatted using Equation 5.3 (Papageorgiou, 2009) to express the percentage degree of severity (PDS) of malaria outbreak.

$$PDS = \begin{cases} 0, & x \leq 0.5 \\ \frac{x-0.5}{0.5} \times 100\%, & x > 0.5 \end{cases} \quad \text{Equation 5.3}$$

where x is the value of the malaria concept from the final vector, Vf . The value will be between 0 and 1, whereby 0 indicates that the malaria concept is not present (0%), and 1 (100%) indicates that the malaria concept is fully present. From 0 to 1 will indicate various degrees of malaria concept presence. Equation 5.3 was used for malaria outbreak interpretation because of its simplicity and a record of successful use in disease predictions (Papageorgiou *et al.*, 2009; Hoyos, Aguilar and Toro, 2022).

For the autumn input concept vector, the malaria outbreak concept showed a value of 0.7856 from the final vector, which, using the above formula, corresponds to 57.2% severity of malaria outbreak. These results are further discretized into four categories

to express malaria outbreak in a human readable format using linguistic variables as illustrated in Table 5.19.

Table 5.19: Linguistic Variables and Corresponding Malaria Incidence Ranges Based on PDS.

PDS	Linguistic variable	Malaria incidences
$0\% < \text{PDS} \leq 30\%$	No-low	0-400
$30\% < \text{PDS} \leq 60\%$	Low	400-800
$60\% < \text{PDS} \leq 70\%$	Moderate	800-1200
$70\% < \text{high} \leq 75\%$	High	1200-1800
$75\% <$	Extreme	>1800

The results were further expressed in terms of malaria classes showing estimation of malaria incidences expected. For example, if the malaria percentage degree severity is between 0% and 25%, very low malaria can be expected with incidences ranging from 0 to 400. Based on the PDS (57.2%) for malaria IKIs reported by the IK expert for autumn, low malaria ranging from 400 to 800 could be expected. The FCM models built in this research study thus provide seasonal malaria predictions.

5.9 Conclusion

This chapter has detailed the construction and utilisation of FCMs to predict malaria outbreaks in the Vhembe district. Using weighted adjacency matrices and FCM Expert software, the complex relationships between various concepts indicative of malaria risk were modelled. These concepts, seasonal weather patterns, environmental indicators plus plants and animal behaviour were visually represented through nodes and their interrelationship through arcs in FCMs for the different seasons: autumn, summer, winter and spring.

The analysis of FCM indices, such as in-degree, out-degree and centrality, provided insights into the relative importance of each concept in the system. In the autumn FCM, "autumn heavy rainfalls" and "dirty water in containers/small pools" were identified as central concepts, highlighting their significance in predicting malaria. The summer FCM underscored the importance of "summer heavy rainfalls," "dirty water in containers" and "summer temperature" as key indicators. On the other hand, the main predictors of malaria in winter were 'Fig "Muhuyu" trees not shedding leaves' and 'sight

of insects/locusts', while spring rainfalls and "mofafa grass" having many ticks' were found important for malaria prediction in spring.

The inference process allowed for the simulation of malaria outbreak scenarios based on observed values of IK Indicators. By employing a sigmoid activation function, the model simulated the spread of influence among concepts until reaching a steady state.

The subsequent chapter continues to focus on malaria outbreak prediction but uses machine learning, weather data and drought indices.

6. Chapter Six: Malaria Prediction Using Weather and Malaria Data

6.1 Introduction

This chapter extends the work of the previous chapter by exploring the prediction of malaria outbreaks using a modern scientific approach. It employs structured historical weather parameters and malaria incidence data to predict malaria outbreaks in the Vhembe district. The chapter explores regression and classification machine learning models to identify the most effective model for forecasting malaria outbreaks in the study area. The discussion begins with a description of the dataset in Section 6.2. After this Section 6.3 addresses the data pre-processing techniques employed, while Section 6.4 focuses on feature selection. Section 6.5 covers the building, training and testing of regression models, including the performance evaluation. Section 6.6 parallels this process by examining classification models and their performance. The chapter concludes with a summary in Section 6.7.

6.2 Dataset

The scientific approach uses machine learning (ML) models that were trained and tested on two datasets from the Vhembe districts. The dataset comprises continuous variables that are heterogeneous because of their diverse sources. The first dataset comprises historical monthly malaria incidence, while the second dataset encompasses the following monthly weather data: precipitation (mm), relative humidity (%), temperature ($^{\circ}\text{C}$). Hydrological drought indices, Effective Drought Index (EDI), and Available Water Resources Index (AWRI) were computed from daily precipitation. Precipitation data were gathered from two climatic districts, rainfall stations 35 and 50, as classified by the South African Weather Services (SAWS). Temperature and relative humidity data were obtained from the Thohoyandou weather station. SAWS operates weather stations to collect parameters such as temperature and relative humidity and the climatic districts refer to the rainfall stations for collecting rainfall (South Africa Weather Service, 2022). The data span 21 years, from January 1998 to December 2018.

The dataset comprised 252 monthly records and included six variables: precipitation, AWRI, EDI, relative humidity, temperature and malaria incidence. Precipitation, AWRI,

and EDI data were collected from weather districts 35 and 50 and aggregated to form single variables. Specifically, AWRI and EDI were averaged, while precipitation was summed. For example, the EDI variable represents the average EDI values from districts 35 and 50. Figure 6.1 illustrates the dataset.

	Vhembe	AWRI	EDI	rainfall	Temp	humidity
Date						
1998-01-31	244	182.40	-0.80	199.5	24.667608	79.629032
1998-02-28	562	141.00	-1.19	23.4	24.531596	76.565287
1998-03-31	312	121.65	-1.33	31.2	25.079704	69.357527
1998-04-30	269	98.65	-1.45	9.2	21.465139	74.320833
1998-05-31	101	77.20	-1.49	0.0	18.835215	57.954301
...	...	...	...	...	...	...
2018-01-08	25	96.95	0.91	11.1	18.614785	53.456989
2018-01-09	336	78.45	0.45	12.2	20.807244	57.401060
2018-01-10	173	98.65	0.58	86.2	22.856164	57.849315
2018-01-11	89	104.95	-0.40	79.0	22.386581	57.806020
2018-01-12	128	115.50	-0.50	91.5	20.303774	66.555256

252 rows × 6 columns

Figure 6.1: A sample of the dataset showing all the variables.

The objective is to identify a function that accurately predicts malaria based on the given features (independent variables). Since the dataset includes the target variable (malaria), this prediction task constitutes a supervised learning problem (Wallisch *et al.*, 2022). Additionally, malaria incidence is represented by continuous values, which called for regression models (Lee, 2022).

Figure 6.2 shows the methodology taken to build ML regression models that predict malaria incidences from pre-processing data to suggest the optimal model for predicting malaria outbreaks. This approach was informed by the Cross-Industry Standard Process for Data Mining (CRISP) methodology (Studer *et al.*, 2021).

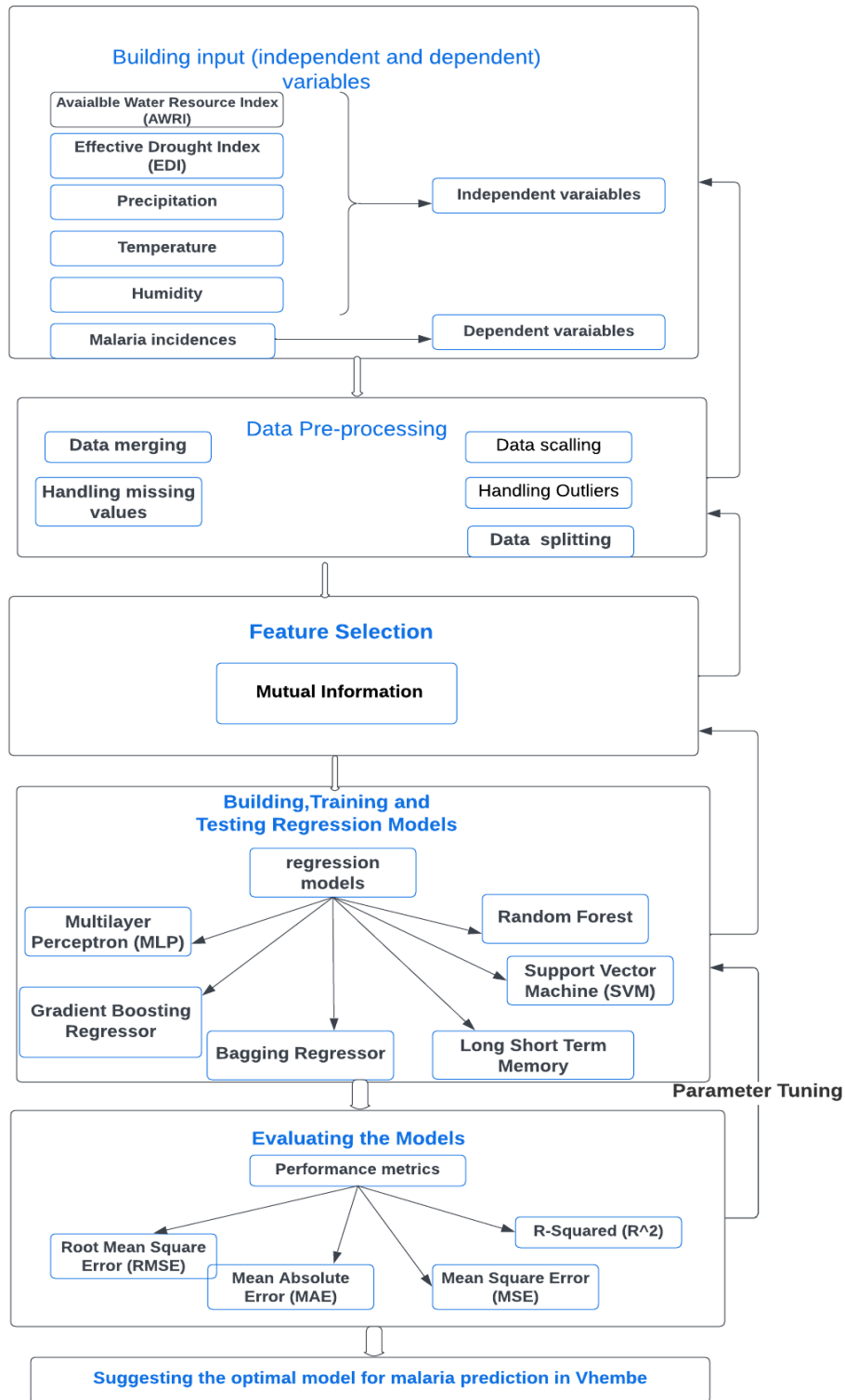


Figure 6.2: Methodology taken to build ML regression models to predict malaria incidences.

6.3 Data Pre-processing

Scikit-learn and Pandas, Matplotlib and NumPy libraries in Python version 3.10.4 were used to create predictive models, process and visualise data.

6.3.1 Data Cleaning

Before the data was fitted into a model, it was pre-processed to ensure it was in the right and appropriate format for the model to use (Jebli *et al.*, 2021). The subsequent section discusses data cleaning procedures undertaken in this study to prepare the dataset for ML regression models.

6.3.2 Merging Datasets of Different Time Series

The sourced dataset presented several challenges. One of the challenges was that the dataset consisted of variables with differing time series. While temperature, relative humidity and rainfall data were recorded hourly, malaria incidence data were provided as monthly figures. To align the weather data with the malaria dataset, the monthly averages of hourly temperature and relative humidity data and the total monthly rainfall were computed using Python. Additionally, the AWRI and EDI, which were calculated from hourly rainfall data, were also averaged to convert them into monthly data.

6.3.3 Missing Values

The dataset also had a few missing data points. Several methods can be used to handle missing values. The common and simplest method is dropping all entries for the independent variable with missing points (Emmanuel *et al.*, 2021). However, this method was not desirable in this research study because there were few missing data points and discarding the entire data series was not desirable (Little and Rubin, 2019). The other way to handle missing values is by patching the missing values. Some ways to patch data include using the mode, mean or zeros to fill in the missing values (Emmanuel *et al.*, 2021). Patching with zeros was not ideal for this research study because the study deals with temporal data. For example, a missing temperature is different from 0°C. On the other hand, interpolation is appropriate for patching the missing values by estimating the best value to fill the missing data points (Caruso, 1998). Caruso asserts that interpolation can be done, for example, by using the last or next valid values to patch the missing values. Interpolation was found to be appropriate in this research study to address the missing data points in the datasets.

In this study, only two independent variables, temperature and relative humidity, had missing data points. Only 1.6% of temperature (four entries) and 6% (15 entries) relative humidity data were missing. Since the percentage of missing data was low, none of these independent variables were excluded from the data sample. There was also a missing value, a single data point, in the dependent variables: malaria incidences.

Interpolation was used to address the missing values in the dataset. Initially, a new data frame was created from the temperature, relative humidity and malaria columns that contained no missing values. A copy of this data frame was then made to include several missing values. Eight interpolation techniques were used to estimate the missing values. Error scores for each technique were calculated by comparing the estimated values against the actual values. The most effective technique for the temperature, relative humidity and malaria incidence variables was Linear Interpolation, `InterpolateLinear()`. Consequently, this technique was used to address the missing data points for temperature, relative humidity and malaria data. Figure 6.1 below illustrates the results of the tests conducted to determine the optimal interpolation method for filling missing data points, which shows Linear Interpolation achieving the highest performance with an R-squared value of 0.9989449630895052.

Table 6. 1: Performance of interpolation techniques for patching missing data points.

Ranking	Method	R-Squared (R^2)
1	InterpolateLinear	0,9989449630
2	FillMedian	0,9974413797
3	InterpolateSLinear	0,9974413797
4	InterpolateAkima	0,9965632381
5	InterpolateCubic	0,9895539332
6	InterpolateQuadratic	0,9888442734
7	InterpolatePoly5	0,2109421173
8	InterpolatePoly7	0,2053645814

6.3.4 Outliers

Outliers were identified in the dataset, as shown in Figure 6.3, and these are values that fall significantly outside the typical range of data points (Sullivan, Warkentin and Wallace, 2021). These values lie 1.5 times the interquartile range below the first quartile (Q1) or 1.5 times the interquartile range above the third quartile (Q3). They

can negatively impact the accuracy of machine learning models by obscuring patterns in the data (Chen and Bien, 2020).



Figure 6.3: Boxplot indicating the presence of outliers in the dataset.

The Interquartile Range (IQR) method was also used to detect these outliers, as depicted by the Python code in Figure 6.4.

```
In [13]: def find_outliers_IQR(malaria_df):
          q1=malaria_df.quantile(0.25)
          q3=malaria_df.quantile(0.75)
          IQR=q3-q1
          outliers = malaria_df[((malaria_df<(q1-1.5*IQR)) | (malaria_df>(q3+1.5*IQR)))]
          return outliers

In [14]: outliers=find_outliers_IQR(malaria_df)

In [15]: outliers
```

Figure 6.4: Python code for identifying outliers and non-outliers in the dataset.

Figure 6.5 below shows some data points in the dataset that have been detected as outliers.

	Vhembe	AWRI	EDI	rainfall	Temp	humidity
Date						
1998-01-31	NaN	NaN	NaN	NaN	NaN	NaN
1998-02-28	NaN	NaN	NaN	NaN	NaN	NaN
1998-03-31	NaN	NaN	NaN	NaN	NaN	NaN
1998-04-30	NaN	NaN	NaN	NaN	NaN	NaN
1998-05-31	NaN	NaN	NaN	NaN	NaN	NaN
...	...	...	...	...	...	...
2017-08-31	NaN	NaN	NaN	NaN	NaN	NaN
2017-09-30	1679.0	NaN	NaN	NaN	NaN	NaN
2017-10-31	1639.0	NaN	NaN	NaN	NaN	NaN
2017-11-30	NaN	NaN	NaN	NaN	NaN	NaN
2017-12-31	1225.0	NaN	NaN	NaN	NaN	NaN

240 rows × 6 columns

Figure 6.5: A view of some outliers in the dataset.

The 'NaN' indicates non outliers. For example, the malaria value for 30/09/2017 (1679) was one of the values identified as an outlier. The outliers were removed and replaced with the median of the dataset, as depicted in the code in Figure 6.6 (Maniruzzaman *et al.*, 2018).

```
In [42]: def non_outliers_IQR(malaria_df):
          q1=malaria_df.quantile(0.25)
          q3=malaria_df.quantile(0.75)
          IQR=q3-q1
          non_outliers = malaria_df[~((malaria_df<(q1-1.5*IQR)) | (malaria_df>(q3+1.5*IQR)))]
          return non_outliers

In [43]: malaria_df=non_outliers_IQR(malaria_df)

In [44]: malaria_df = malaria_df.fillna(malaria_df.median(axis = 0, skipna = True))
```

Figure 6.6: Addressing outliers using the median.

6.3.5 Data Normalisation

The sourced dataset had features with varying ranges of dimensions. For example, the EDI values range from -2 to 4, while features such as AWRI and malaria incidences were going into the hundreds. The standard scaler function, `StandardScaler()`, from scikit-learn library in Python was used to normalize all the series in the dataset to the same scale. This was done because neural networks such as multilayer perceptron (MLP), one of the models considered in this research study, are sensitive to the scale of the features. Figure 6.7 illustrates the code that uses the `StandardScaler()` function to scale the dataset; x , in the code, represents the dataset.

```
#Function to scale the data using standered scaler
def apply_standard_scaler(x):
    scaler = StandardScaler()
    scaler.fit(x)

    return scaler.transform(x)
```

Figure 6.7: `StandardScaler()` function to scale the dataset.

6.4 Feature Analysis

Features (input variables) may not be equally important in predicting the target output variable, malaria (Zhou *et al.*, 2022). A feature analysis was conducted to determine which features have the strongest relationship with the target variable, malaria. This process also allowed for selecting only features most relevant to predicting malaria outbreaks (Zhou *et al.*, 2022). Additionally, feature selection was performed to minimise computational costs and enhance the performance of the models (García, Luengo and Herrera, 2015).

Mutual Information

Mutual information (MI) is a metric that quantifies the extent to which one random variable provides information about another. It represents the reduction in uncertainty about one random variable when the value of another is known. This research study used MI to quantify the relationship between each input variable and the target variable. Table 6.2 below illustrates the contribution of each input variable to the target variable, malaria.

Table 6.2: Mutual information for features in the prediction of malaria in Vhembe District.

Feature/input variable	Mutual Information
AWRI	0,9231
EDI	0,7249
Rainfall	0,0000
temperature	0,4627
Relative Humidity	0,0000

Table 6.2 presents the mutual information values between input and target variables. AWRI exhibits the highest mutual information (0,9231), indicating a strong dependency on the target variable. EDI follows with a value of 0.7249, suggesting a moderate relationship. Temperature, with a mutual information value of 0.4627, also shows a meaningful, though weaker, association with the target. In contrast, rainfall and relative humidity have mutual information values of 0.0000, indicating no detectable dependency on the target variable. This suggests that AWRI, EDI and temperature are the most informative predictors in this dataset.

Lag Determination

Lag testing was performed to determine the appropriate time lag between weather conditions, hydrological drought indices and malaria outbreaks. A time lag exists between the growth of the *Anopheles* mosquito, the vector for malaria, from the egg to the onset of symptoms in humans (Segun *et al.*, 2020). Incorporating this lag into the model can enhance its accuracy in predicting how these weather conditions and drought indices influence malaria outbreaks. Time lags of one, two, three and four months were tested for all features. Table 6.3 presents the time lags that yielded the optimal mutual information for each feature. AWRI exhibited perfect mutual information, indicating its significant impact on malaria outbreaks.

Table 6.3: Mutual Information after the optimum lag times.

Feature/input variable	Mutual Information	Lag time (months)
AWRI	1,0000	1
EDI	0.9624	1
Rainfall	0.8000	2
Temperature	0,5275	1
Relative Humidity	0.5234	2

AWRI, EDI and temperature showed the highest MI score to the target variable at a time lag of one. On the other hand, relative humidity and rainfall had optimum MI at the time lag of two months. Since all the features have an MI score greater than 0,5, all features were selected for use in the prediction of the target variable, malaria outbreak.

After data pre-processing, the dataset was divided into training and testing subsets with a ratio of 8:2 (Nguyen *et al.*, 2021). A larger training dataset is essential to provide the model with sufficient information to learn the patterns within the data effectively and make the models more stable (Tham *et al.*, 2020).

6.5 Regression Models Results

The following machine learning models were trained and tested to determine the optimal model for malaria outbreak prediction: Multi-layer Perceptron (MLP) (artificial Neuron network (ANN)), Random Forest (RF), Support Vector Machine (SVM), Decision Trees, Long Short-Term Memory (LSTM), Deep Learning, (DP), gradient Boosting and bagging models. The literature review showed that these supervised machine learning algorithms have successfully been used to predict malaria outbreaks.

The performance metrics discussed in Chapter 3 were used to evaluate the models' predictive performance. The metrics evaluate the models by comparing the simulated values during the training and testing process against the actual values. Table 6.4 presents the performance of the models.

Table 6.4: Performance of the developed models.

Models	R ²	MSE	RMSE	MAE
MLP				
Decision Trees	0.09	74953.24	273.78	188.71
Bagging Regressor	0.53	3646.93	60.39	41.91
SVM	0.49	88610.07	297.67	203.40
Random Forest	0.38	3646.93	60.39	41.92
LSTM	0.39	0.73	0.85	0.64
Gradient Boosting Regressor	0.43	47264	217	166
Deep Learning	0.41	0.35	0.59	0.39

It can be noted that the models did not show satisfactory performance that can be reliable for malaria prediction, meaning a reliable function that accurately predicts

malaria based on the given features. The highest R^2 accuracy of 53% was achieved by the Bagging Regressor model, albeit with a high MSE, RMSE and MAE compared to the other models. On the other hand, LSTM showed lower errors for MSE, RMSE and MAE, but the R^2 value was only 39%. It is important to note that the results were best achieved after parameter tuning, which was performed to address the low performance of the models.

The regression models yielded low performance in predicting malaria outbreaks necessitating the exploration of classification models. Classification models are also well-suited for predicting categorical outcomes and can provide accurate predictions when dealing with discrete events (Vujovic, 2021). Classification models can model malaria outbreaks by categorising data into distinct classes based on predictive features and target variables (malaria). This shift in methodology aims to develop a more robust model which can improve the accuracy and reliability of MOEWS.

6.6 Classification Models

Due to their widespread use and success in modelling classification tasks, the following classification models were considered in this study: MLP classification, Decision Trees classification, Naïve Bayes Classification, AdaBoost classification and SVM classification, as well as K-Nearest Neighbors (KNN) algorithms (Kalipe, Gautham and Behera, 2018; Bahl *et al.*, 2019; Abdullah and Abdulazeez, 2021; Mohapatra *et al.*, 2022). Figure 6.8 illustrates the flowchart of the approach to building the classification model for predicting malaria outbreaks.

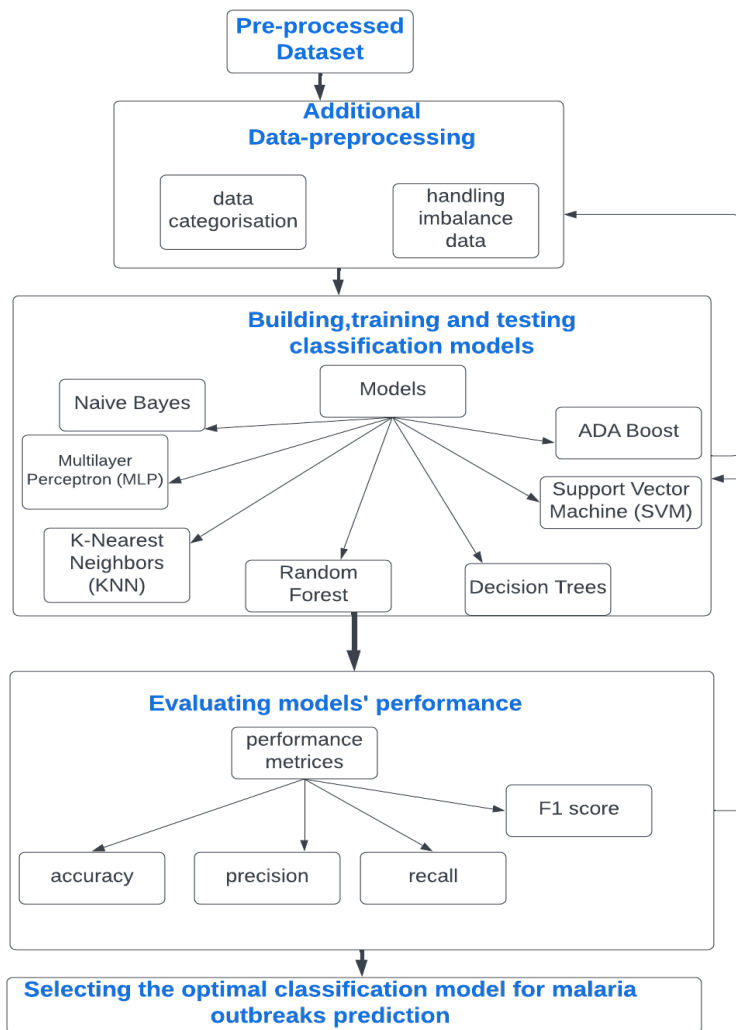


Figure 6.8: The process of identifying an optimal classification model.

6.6.1 Converting Continuous Variables to Categorical Variables

Most data cleaning processes, including addressing missing values and outliers, were already performed when preparing the dataset for regression prediction. Therefore, in the classification prediction, the initial step was to convert all variables, both features and the target variable, into categorical values. Firstly, the maximum value in each variable was determined and then divided by 5 to get the dimension of the five segments or classes. Equation 6.1 illustrates finding the dimension (size, x) of each segment for a variable, based on max , the maximum value of the original continuous variable:

$$Px = \frac{max}{5} \quad \text{Equation 6.1}$$

The interval $[0, max]$ is divided into five segments, each with a dimension of x , and these segments are labelled with categorical values of 1, 2, 3, 4 and 5. These categorical values refer to classes. The next step was to compute the categorical value for each continuous value using Equation 6.2.

$$categoricalValue = \left\lfloor \frac{value-1}{x} \right\rfloor \quad \text{Equation 6.2}$$

whereby *value* refers to a continuous value that is being converted into a class/categorical value, while x is the segment size for the particular variable, as obtained from Equation 6.1 above. For example, to categorise relative humidity value of 33, since relative humidity ranges from 0 to 100%, it was segmented into 20 equal parts using Formula 6.1 as shown below:

$$x = \frac{100}{5} = 20$$

The categorical value of 33 was computed by substituting values in Formula 6.2 as follows:

$$categoricalValue = \left\lfloor \frac{33-1}{20} \right\rfloor + 1 = 2$$

Figure 6.9 below depicts the dataset containing the categorical data

	Malaria	AWRI	EDI	Rainfall	Temperature	Humidity
Date						
1998-01-31 00:00:00	1	2	4	1	4	4
1998-02-28 00:00:00	2	2	4	1	4	3
1998-03-31 00:00:00	1	2	4	1	5	3
1998-04-30 00:00:00	1	1	4	1	4	4
1998-05-31 00:00:00	1	1	3	1	3	3
...	...	...	...	...	...	...
2018-08-31 00:00:00	1	1	4	1	3	3
2018-09-30 00:00:00	1	1	4	1	4	3
2018-10-31 00:00:00	1	1	4	1	4	3
2018-12-31 00:00:00	1	1	4	1	4	3
31/11/2018 00:00:01	1	1	4	1	5	4

252 rows × 6 columns

Figure 6.9: dataset showing categorical.

Table 6.5 below shows the malaria classes as defined by Equation 6.1 and their corresponding incidence ranges. Linguistic variables are also provided for easier interpretation.

Table 6. 5:Malaria classes and their corresponding malaria incidences brackets with linguistic representation.

Malaria Class	Malaria incidences	Linguistic variable
1	0-400	No-low
2	400-800	Low
3	800-1200	Moderate
4	1200-1800	High
5	>1800	Extreme

6.6.2 Handling Imbalanced Data

The primary challenge associated with the malaria prediction dataset for classification purposes was its substantial imbalance. The dataset comprised 252 rows, with the distribution of instances as follows: 188 rows for class 1 with far fewer, 41, 13, 5 and 5 rows for classes 2, 3, 4, and 5, respectively. Figure 6.10 shows the imbalance within the dataset.

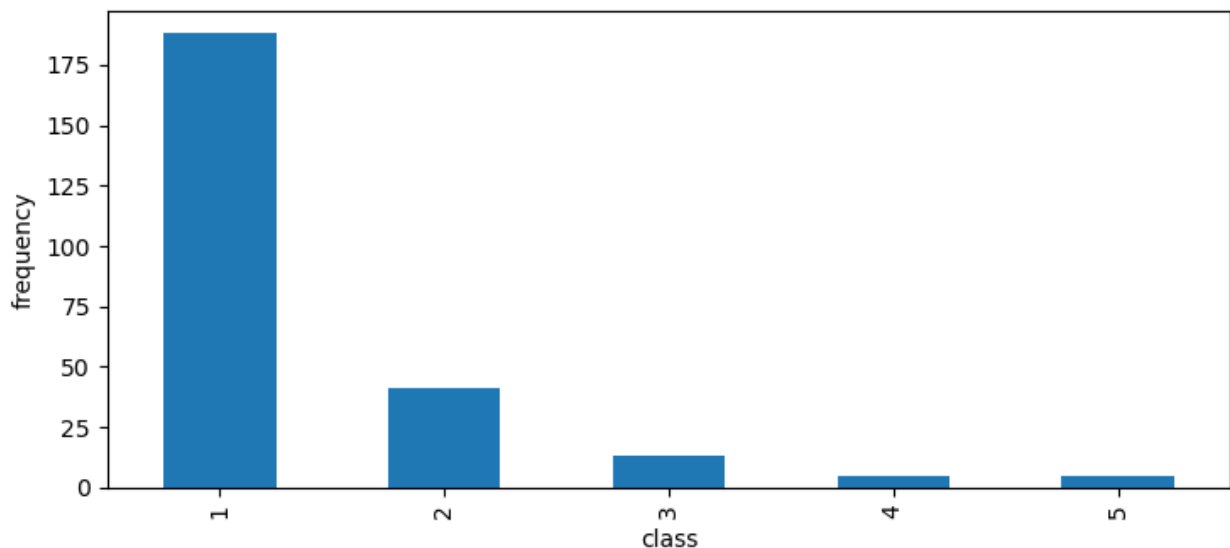


Figure 6. 10: Imbalanced dataset.

We refrained from training the machine learning models on the highly imbalanced dataset, as doing so could result in inflated accuracy metrics while potentially compromising precision and recall (Sailasya and Kumari, 2021). Without addressing

this imbalance, the outcomes may be unreliable and the predictions may lack accuracy. We addressed this issue by using oversampling, a technique that equalises the class distribution by increasing the representation of minority classes to match the size of the majority class (García *et al.*, 2020). Figure 6.11 provides Python code used for oversampling.

```
: #oversampling the dataset to address imbalance in the dataset
from sklearn.utils import resample
#create different dataframe , one of majority and 4 minority classes
df_majority = malaria_df[(malaria_df['Malaria']==1)]
df_minority2 = malaria_df[(malaria_df['Malaria']==2)]
df_minority3 = malaria_df[(malaria_df['Malaria']==3)]
df_minority4 = malaria_df[(malaria_df['Malaria']==4)]
df_minority5 = malaria_df[(malaria_df['Malaria']==5)]

# upsample minority classes
df_minority2_upsampled = resample(df_minority2,
                                replace=True, # sample with replacement
                                n_samples= 188, # to match majority class
                                random_state=42) # reproducible results

df_minority3_upsampled = resample(df_minority3,
                                replace=True,
                                n_samples= 188,
                                random_state=42)

df_minority4_upsampled = resample(df_minority4,
                                replace=True,
                                n_samples= 188,
                                random_state=42)

df_minority5_upsampled = resample(df_minority5,
                                replace=True,
                                n_samples= 188,
                                random_state=42)

# Combine majority class with upsampled minority class
df_upsampled = pd.concat([df_minority2_upsampled, df_minority3_upsampled, df_minority4_upsampled,
                        df_minority5_upsampled, df_majority])
```

Figure 6.11: Code snippet for oversampling data.

The oversampling process yielded a dataset comprising 912 records. Figure 6.12 displays the distribution of malaria classes following the oversampling procedure.

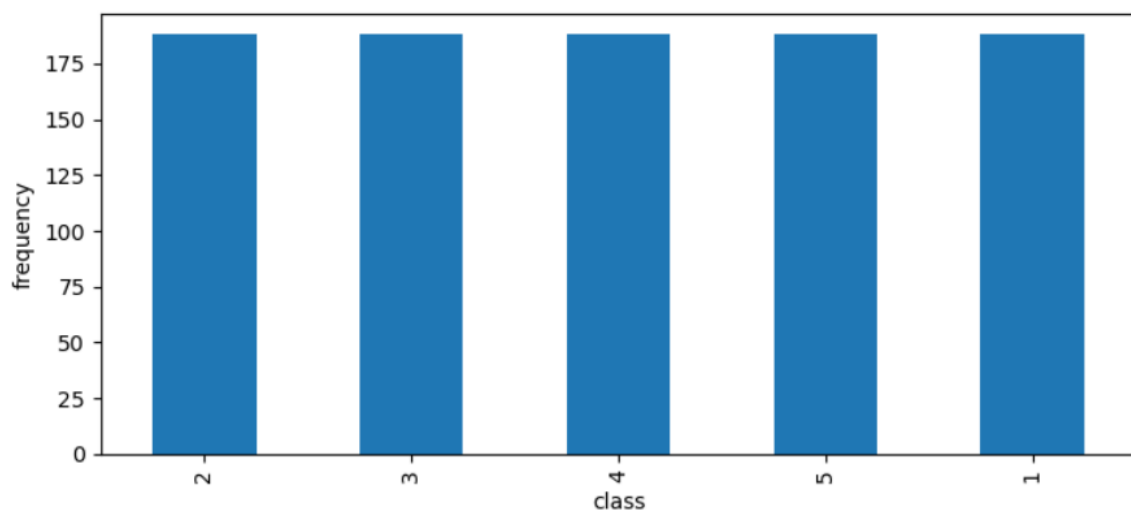


Figure 6.12: Distribution of the malaria categories/classes after oversampling.

After cleaning the data, the dataset was split into training and testing, whereby 80% of the data was used to train the models while the remaining 20% was used for testing purposes. The results obtained during the testing from all the models are discussed in the next section.

6.6.3 Classification Models Results

The models were compared in terms of their performance using the F1-score accuracy, precision and recall scores. Precision The results are shown in Figures 6.13, 6.14, 6.15 and 6.16, respectively.

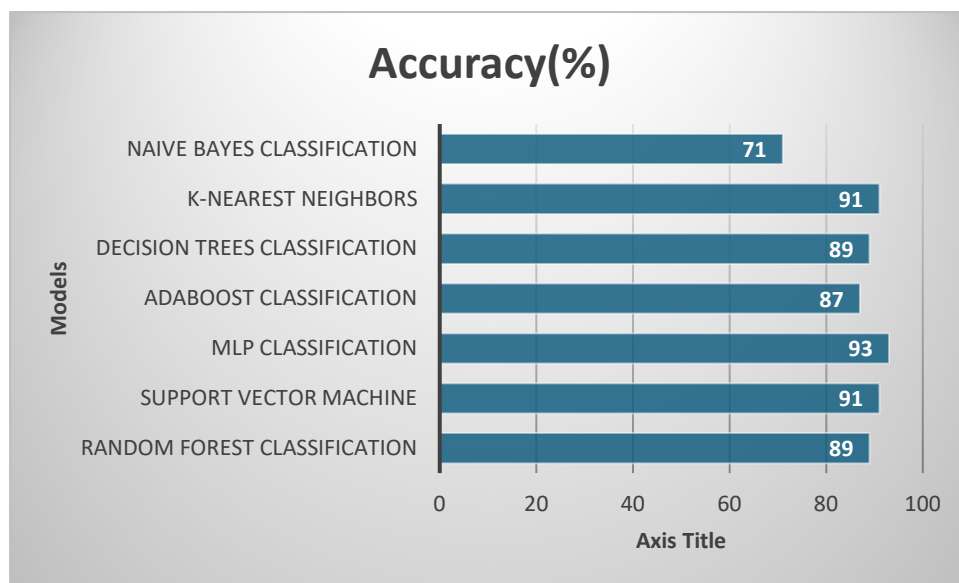


Figure 6.13: Comparing the accuracy score of the models.

Figure 6.13 illustrates the accuracy of various machine learning models in predicting malaria outbreaks using weather parameters. MLP model achieves the highest accuracy at 93%, indicating strong classification capability. MLP was followed by SVM and KNN models, each with an accuracy of 91%. The Random Forest and Decision Trees models both have an accuracy of 89%, showing robust performance. AdaBoost follows with 87%, while the Naive Bayes Classification model has the lowest accuracy at 71%, suggesting it may be less effective for malaria prediction in this study. Overall, the MLP, SVM and KNN models demonstrate high accuracy in forecasting malaria outbreaks based on weather conditions.

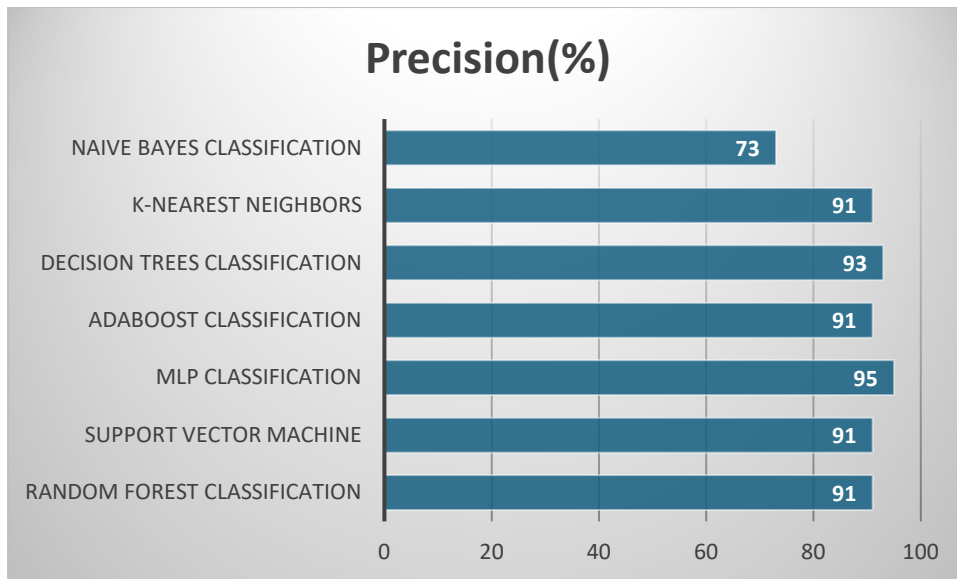


Figure 6.14: Precision score of the models.

Figure 6.14 shows that MLP model outperformed other models with a precision of 95%. Decision Trees follows closely with a precision of 93%. The KNN, AdaBoost, SVM and Random Forest models each have a precision of 91%. On the other hand, the Naive Bayes Classification model shows the lowest precision at 73%, highlighting its relatively weaker performance in correctly predicting malaria outbreaks. Considering different scores, MLP and Decision Trees models exhibit better performance, making them the most accurate in forecasting malaria outbreaks based on weather data.

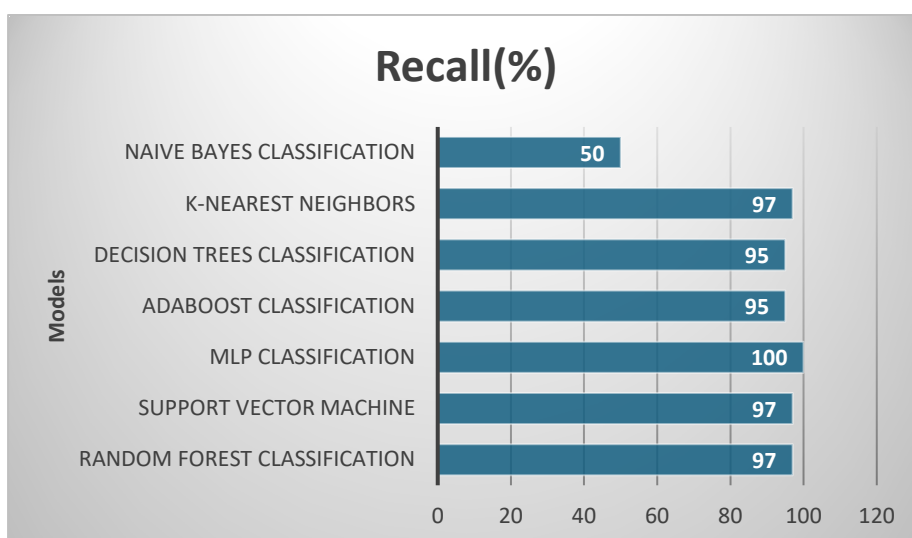


Figure 6.15: Recall score achieved by the models.

The MLP classification model has a perfect recall score, making it the best choice for malaria outbreaks, as capturing all positive instances is critical. In contrast, the Naive Bayes classifier performs poorly with a recall of 50%, indicating a significant risk of missing relevant instances. The other models - Random Forest, SVM, KNN, AdaBoost, and Decision Trees - show competitive recall scores with minor differences, making them suitable for malaria outbreak predictions.

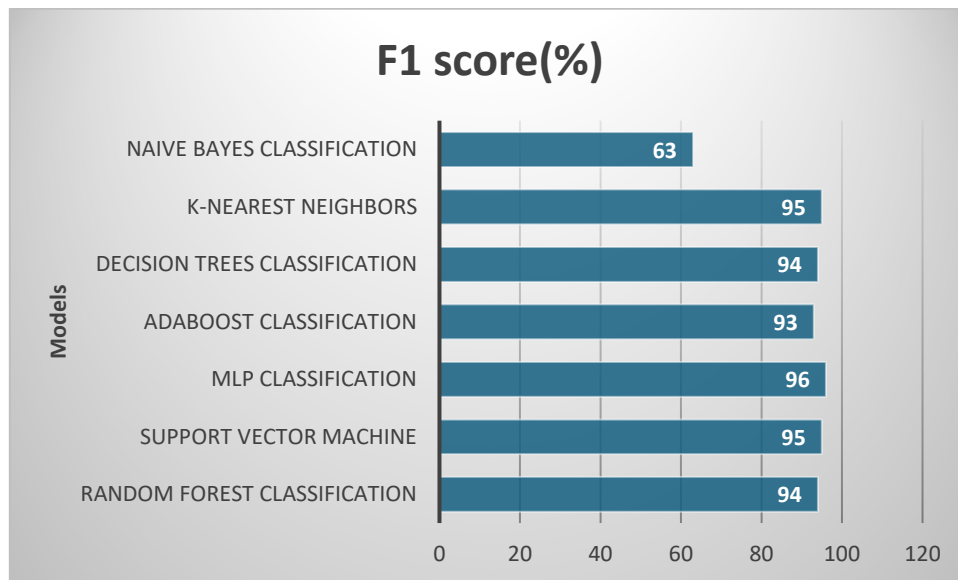


Figure 6.16: F1 score of the models.

The F1 score is an important performance metric as it balances precision and recall, providing a comprehensive measure of a model's ability to predict outbreaks accurately. Among the models analysed, the MLP classification model demonstrates the highest F1 score of 96% (as seen in Figure 6.16), making it the most effective in balancing the accurate prediction of malaria outbreaks while minimizing false alarms. SVM and KNN also perform well, each with an F1 score of 95%. Random Forest and Decision Trees follow closely, each with an F1 score of 94%. AdaBoost classification, with an F1 score of 93%, still performs adequately but shows a slightly lower effectiveness in balancing the trade-off between detecting outbreaks and avoiding false positives. In contrast, the Naive Bayes classification model, with a significantly lower F1 score of 63%, seems to provide less reliable predictions in the context of malaria outbreaks, likely due to its limitations in handling the complexity and interdependence of features in such data (Sailasya and Kumari, 2021).

The MLP classification model was chosen for the deployment of malaria prediction in the Vhembe district, considering that it showed the best performance. The model was

deployed to predict malaria from March to October 2023, given the weather parameters and drought indices depicted in Table 6.5 below. Using the procedure described in Section 6.6.1, the data was converted into the categorical format before input into the model for prediction.

Table 6.6: Weather parameters and drought indices used to deploy MLP model for March to October 2023 malaria prediction.

Rainfall	Temperature	Relative Humidity	EDI	AWRI
64,1	21	66	-0.84	126,0
49,5	19	87	-0.67	141,2
8,9	16	70	-1.10	69,2
4,3	15	52	4.30	169,9
7,1	12	54	0.31	103,3
6,5	17	55	1.04	204,3
25,4	19	51	-0.53	175,7
39,4	20	55	-1.93	57,5

Table 6.6 shows malaria incidences recorded from March to October.

Table 6.7: Malaria incidences reported in Vhembe in 2023.

Month	Malaria incidences
March	25
April	14
May	15
June	7
July	9
August	11
September	20
October	20

Similarly, malaria incidences were also categorised to allow comparison of the predicted class with the actual classes. Table 6.7 shows the classes predicted by the model vs the actual classes of the recorded malaria data.

Table 6. 8: Predicted malaria classes vs actual malaria classes.

Month	Predicted Class	Actual class
March	2	1
April	1	1
May	1	1
June	1	1
July	1	1

August	1	1
September	1	1
October	1	1

Figure 6.17 reflects the graphical view of the predicted values against the actual values.

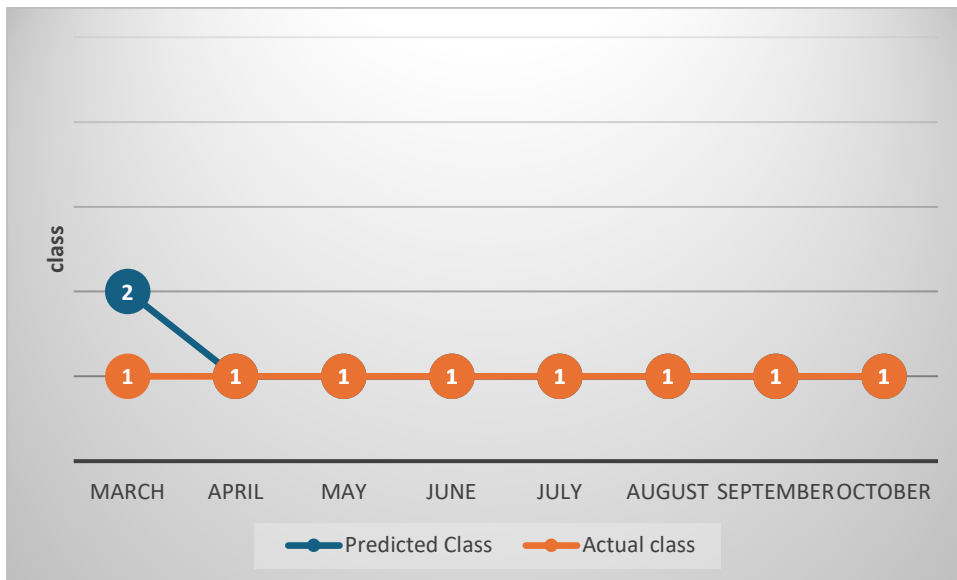


Figure 6. 17: Graphical view predicted classes vs actual classes.

Figure 6.18 shows the confusion matrix from the MLP model

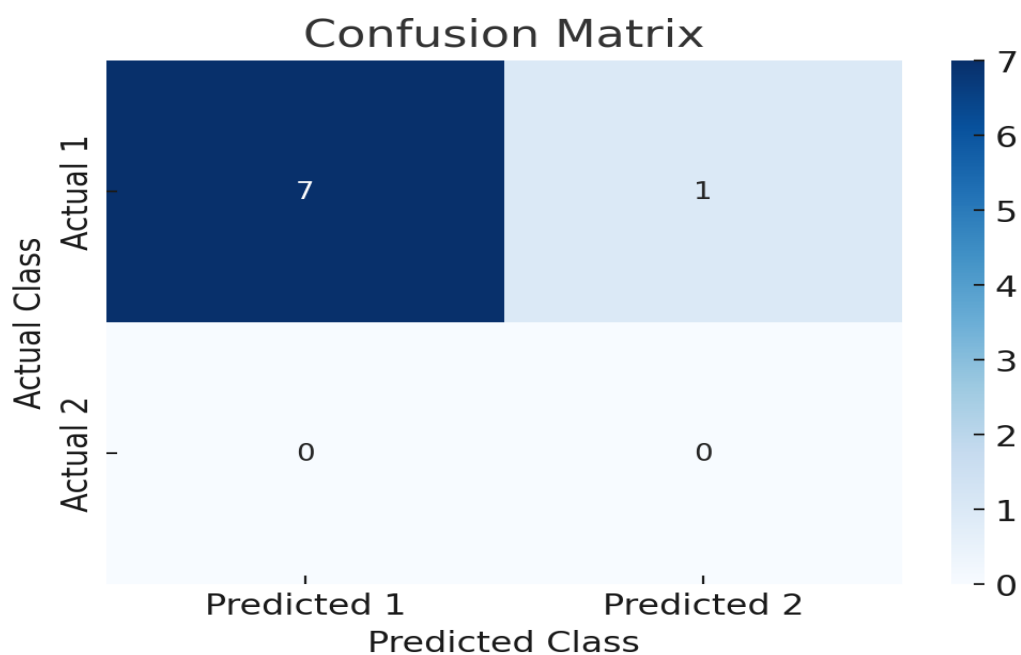


Figure 6. 18: Confusion matrix for MLP during the deployment.

The model has performed well, accurately predicting most of the classes. However, the model predicted a class of 2 instead of 1 in March. These results showed an accuracy of 75%, a precision of 100%, a recall of 75% and F1 of 86%. This suggests there may be a few cases where the model predicts incorrectly, indicating room for further refinement to enhance accuracy.

6.7 Discussion of Results

While our classification models performed well in using weather data to predict malaria outbreaks, our regression models could not accurately capture the relationship between weather parameters and malaria incidences. This underperformance highlights the need for further research to improve the ability of regression models to learn from the data. Classification models for malaria forecasting have been widely used in other studies and have demonstrated strong predictive accuracy (Nkiruka, Prasad and Clement, 2021).

The feature selection process revealed that rainfall strongly correlates with malaria incidences, showing its importance in predicting outbreaks. Rainfall has consistently been identified as a key predictor of malaria in several studies (Makinde and Abiodun, 2020; Tiu *et al.*, 2021). For instance, Githeko *et al.* (2018) concluded that increased rainfall doubled the number of mosquitoes. Our findings in Vhembe align with results obtained by Sehlabana, Maposa and Boateng (2020), who also identified rainfall as a key predictor of malaria in the region. The 2-month lag we observed in rainfall's effect on malaria incidences is consistent with the study by Gonzalez-Daza *et al.* (2023). However, other studies, such as Nanvyat *et al.* (2018), found 3- and 5-month lags in rainfall's effect on malaria in different regions of Nigeria. These differences in time lags are explained by Nkiruka, Prasad and Clement (2021), who note that the impact of climate variability on malaria incidence is not uniform and varies across regions.

Relative humidity and temperature also showed high mutual information values above 0.5, indicating a strong relationship with malaria incidences. These findings align with several other studies (Kim *et al.*, 2019; Nkiruka, Prasad and Clement, 2021; Mohapatra *et al.*, 2022). In our study, we also observed a 2-month lag for relative humidity and a 1-month lag for temperature in relation to malaria incidence, which is consistent with

other research studies. A study by Kumar, Pisudde and Sarthi (2022) concluded that temperature impacts malaria incidences within a 2-month lag.

Our findings confirm the influence of weather parameters on malaria incidences, as reported in several previous studies (Hoshen and Morse, 2004; Goswami *et al.*, 2014; Darkoh, Larbi and Lawer, 2017; Segun *et al.*, 2020; Ebi *et al.*, 2021). However, we also introduce a novel finding: the effect of drought indices on malaria. The high mutual information values indicate that the AWRI and EDI strongly correlate with malaria incidences. The use of EDI and AWRI in predicting malaria has been relatively underexplored. In our study, AWRI emerged as the main predictor of malaria in Vhembe, likely because it effectively measures the environmental conditions conducive to malaria-carrying mosquitoes. However, since a combination of weather parameters influences mosquito breeding conditions, we included all features with mutual information values of 0.5 or higher, as they reflect significant relationships with malaria outbreaks.

6.8 Conclusion

In this chapter, we explored the use of machine learning models for predicting malaria outbreaks in the Vhembe region of Limpopo province, South Africa. Utilising a scientific approach, the datasets containing historical monthly malaria incidence and various weather data and hydrological drought indices were analysed. Intensive data analysis and preparation was carried out. Despite implementing regression models and parameter tuning, the results were unsatisfactory, highlighting the necessity for exploring classification models. Classification models, adept at handling categorical outcomes, provided a more accurate and nuanced understanding of malaria outbreaks. Several classification models were implemented including MLP, Decision Trees, Bayesian Classifiers, SVM, KNN and Random Forest. The imbalanced nature of the dataset required oversampling to ensure accuracy in the predictions. Ultimately, the MLP model outperformed the other models.

The shift in methodology represents a significant step towards enhancing the effectiveness and reliability of MOEWS and underscores the importance of integrating diverse computational techniques to forecast malaria outbreaks.

The next chapter will discuss dissemination and communication of malaria outbreaks informed by the prediction done in the second component of MOEWS, as addressed previously in Chapter 5 and in the current chapter.

7. Chapter Seven: Malaria Outbreaks Communication and Dissemination

7.1 Introduction

This chapter addresses the third component of the Malaria Outbreak Early Warning System (MOEWS), which dwells on disseminating alerts on malaria outbreaks. The chapter aims to enhance the delivery of the warnings to those at risk in an understandable manner using simple and clear information.

The success of early warning systems hinges on their ability to be easily understood, clearly communicated and effectively deliver warnings to those at risk (Masinde, 2012). It is, therefore, crucial that MOEWS meets the requirements of the user, particularly the Vhembe community, who primarily rely on mobile phones for their information and communication technology needs. To increase the likelihood of success for the MOEWS, the system's user interfaces were primarily designed and developed with mobile phones in mind, utilising mobile applications.

A significant contribution of this research was integrating indigenous knowledge (IK) with modern technology to predict malaria outbreaks. The IK indicators and information on malaria risk are made accessible to researchers through a comprehensive web portal. An alert system for extreme weather events, such as heavy floods, was also integrated into the MOEWS.

A mobile phone application was developed for the Android platform to facilitate the collection and storage of IK indicators for malaria outbreaks and reporting extreme weather conditions. A web portal was also developed to disseminate malaria outbreak alerts and an overview of MOEWS. This chapter details the analysis, design, implementation and testing of MOEWS. Section 7.2 discusses the system design, and the Sections 7.3, 7.4 and 7.5 focus on the mobile application, web portal and social media, respectively. Section 7.6 concludes the chapter.

7.2 System Overview and Use Case

Architecture of the System Prototype

Figure 7.1 below illustrates the MOEWS prototype, which aligns with the MOEWS framework, as discussed in Chapter 3.

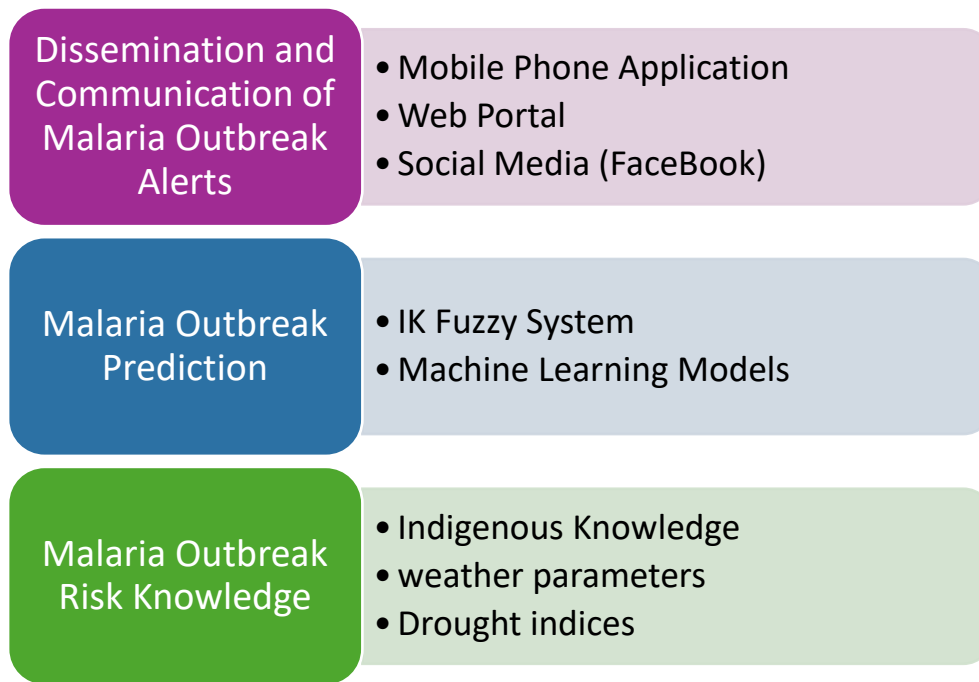


Figure 7.1: MOEWS prototype architecture.

Malaria Outbreak Risk Knowledge

This component is dedicated to collecting the data necessary for forecasting malaria outbreaks. IK data is gathered by the local people using a mobile phone application. Historical weather data is obtained from the South African Weather Service, while current weather data is captured from online published sources and stored in the database. Additionally, drought indices are calculated from the collected rainfall data and stored in the database.

Malaria Outbreak Prediction

This component utilises both modern scientific methods and IK to predict malaria outbreaks. In the realm of modern science, machine learning models are developed, trained and tested on historical weather parameters to produce monthly forecasts, with predictions made one month in advance. These models rely on extensive datasets and advanced algorithms to identify patterns and trends that may indicate potential outbreaks. Conversely, the IK system employs cognitive fuzzy maps and indigenous knowledge indicators to construct a model capable of predicting seasonal malaria outbreaks.

Dissemination and Communication of Malaria Outbreak Alerts

Malaria outbreak predictions generated by the second subsystem are communicated and disseminated to the affected community through a web portal, mobile phone application and the social media platform Facebook. End users must first register to access malaria alerts from the MOEWS mobile phone application. The mobile phone application is used to input and output the malaria outbreak IK indicators and extreme weather conditions. The web portal and Facebook provide comprehensive information on malaria, malaria outbreak alerts and weather. This information is available in various formats, such as text, graphs and audio files.

Figure 7.2 below shows the use case diagram of MOEWS.

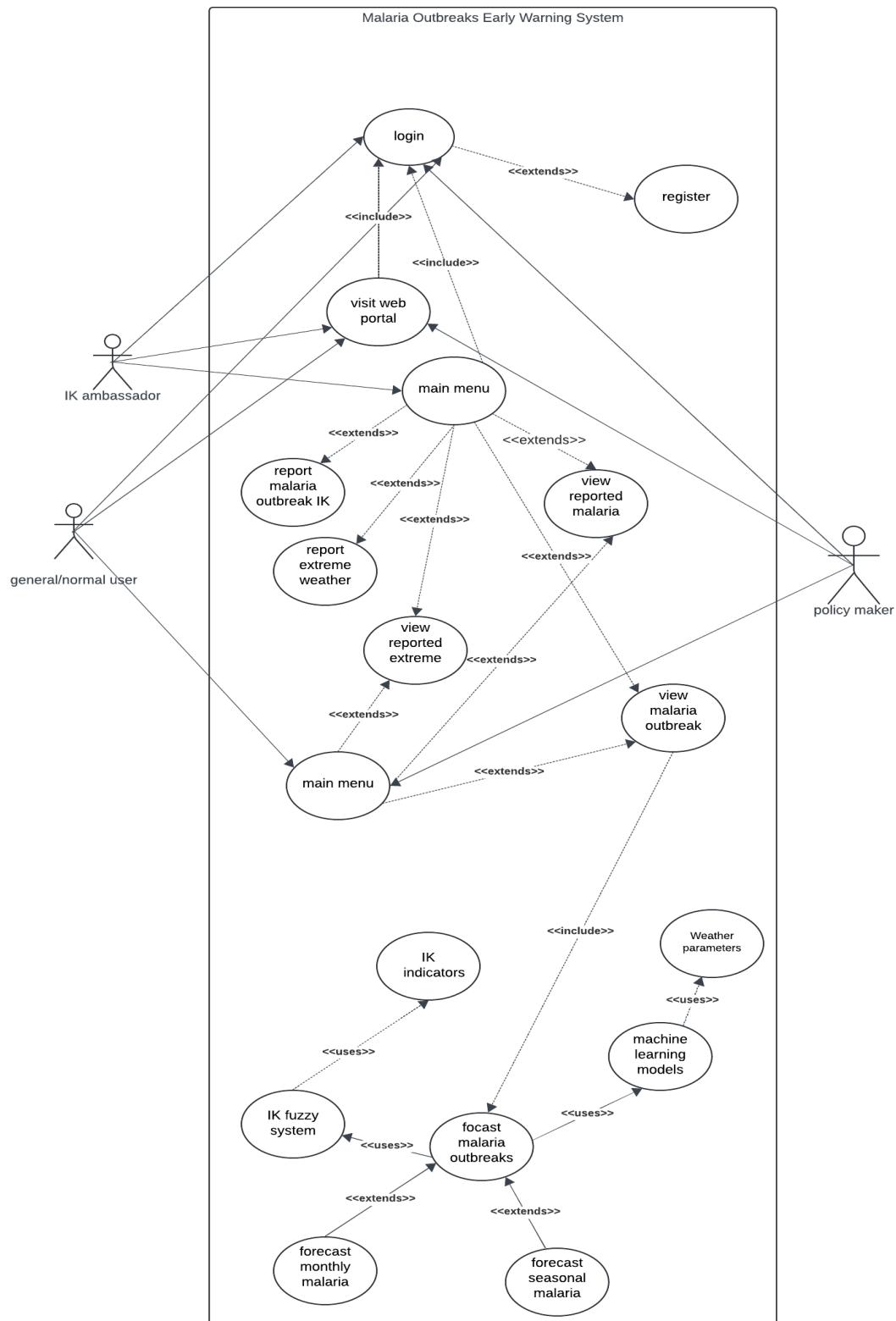


Figure 7.2: MOEWS use case diagram.

Database Design

The system's database was designed using Firebase, incorporating essential integrity constraints as needed. In situations where entity relationships were particularly complex, such as with indigenous knowledge Indicators, stored queries and database triggers were utilised to manage the data effectively. To achieve desirable database features such as faster UPDATE performance, data integrity and query optimisation, the database was normalised. Figure 7.3 shows the entity relationship diagram of the MOEWS database.

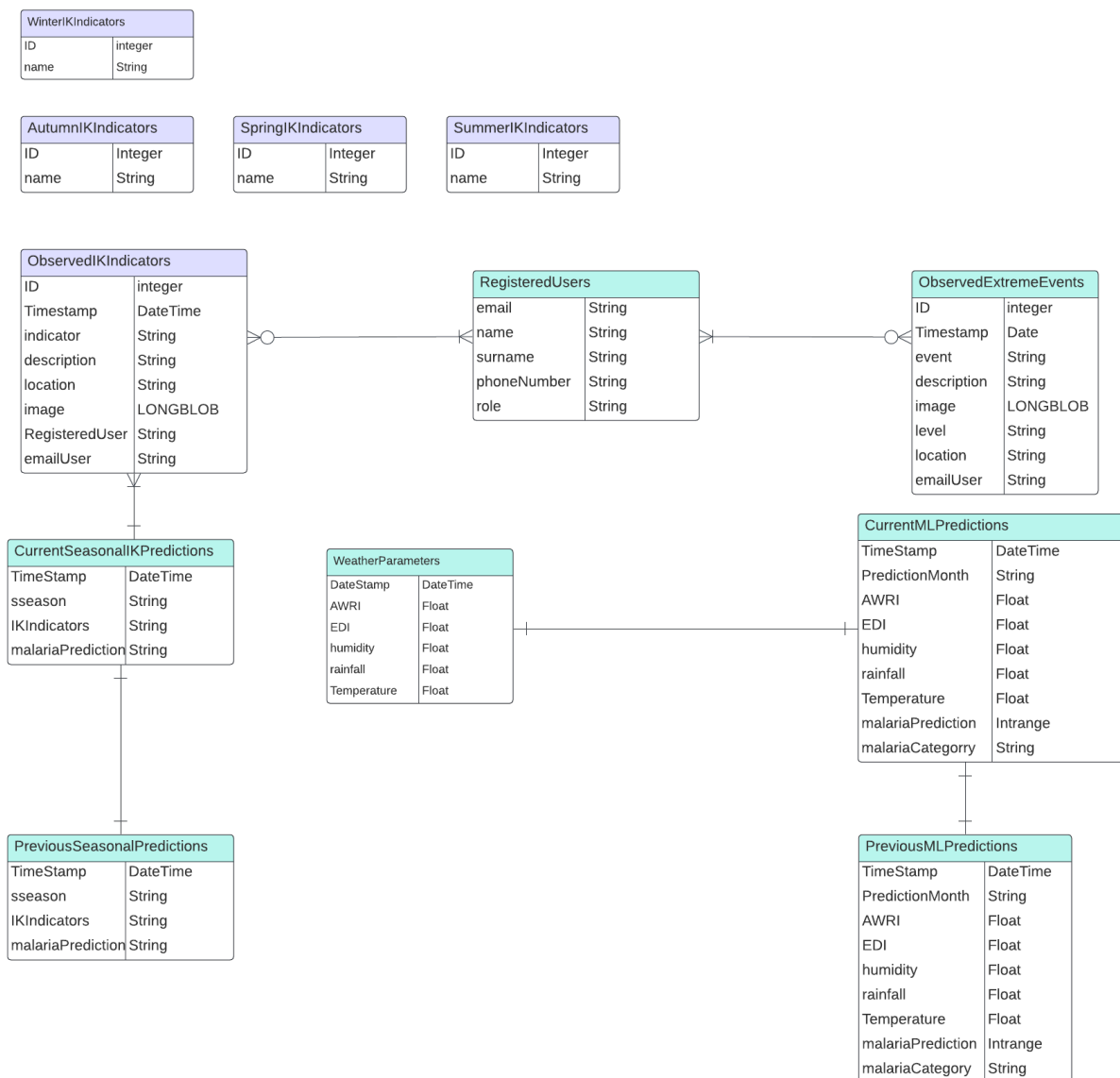


Figure 7.3: MOEWS Entity Relationship Diagram.

7.3 Mobile Phone Applications Implementation

This mobile application is intended for use by IK ambassadors, who act as a link between local communities and the researcher. The application contains pre-defined malaria outbreak indicators stored in a Firebase database. The IK ambassador's task is to record observations based on these indicators. Malaria outbreak indicators are reported by seasons. To account for internet absence, data is stored locally using Firebase's 'persistence' system and synchronised with the remote Firebase database once an internet connection is available. This automatic local storage ensures that IK indicators are cached on the device and uploaded to the central database for storage when connectivity is restored. Considering the link between weather variations and malaria outbreaks, the application also enables the IK ambassador to report observed extreme weather events, such as floods in Vhembe.

The mobile application was developed using Java programming language on Android Studio Flamingo (2022.2.1).

7.3.1 Application Installation

The Android Application Package (APK), a file format used by the Android operating system for the distribution and installation of mobile apps and middleware, was generated from the MOEWS app and used to install the app. MOEWS was installed on a Samsung Galaxy A31, which runs on Android 12 Snow Cone, version 12. Figure 7.4 below shows the MOEWS app appearing under the installed apps.

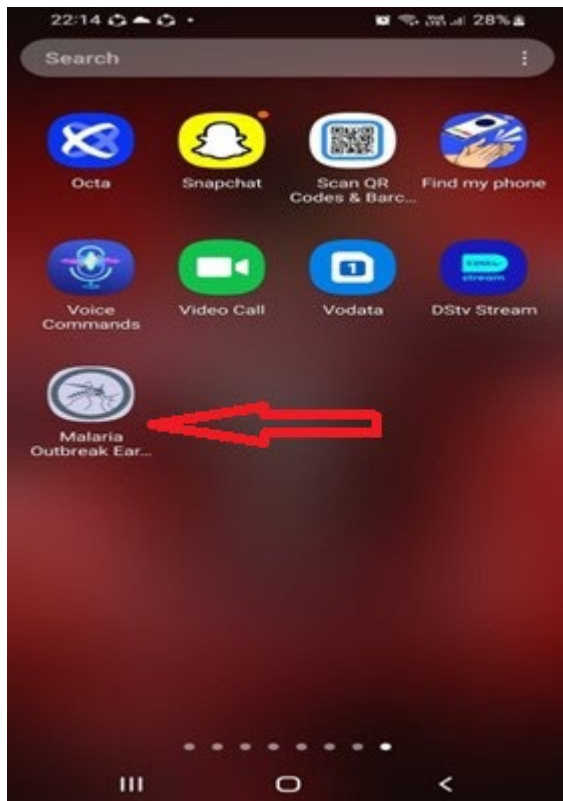
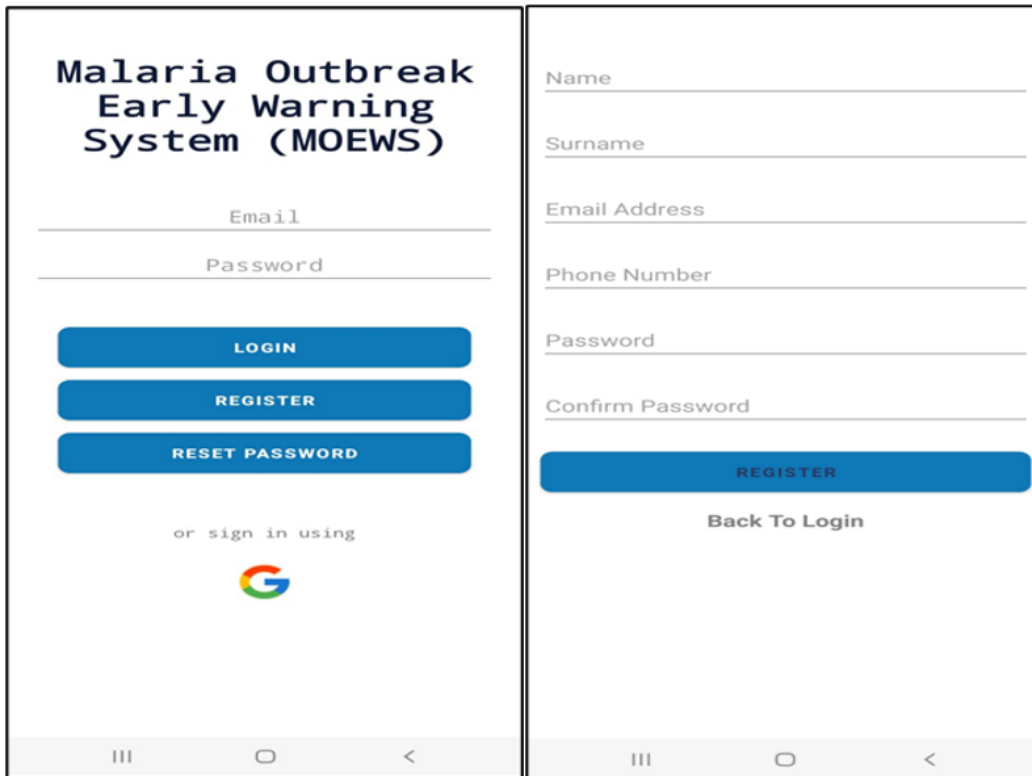


Figure 7.4: icon of MOWES mobile application.

7.3.2 Application Functionalities

Register sub menu

A user needs to register to access the app. Figure 7.5 depicts the registration and login screens, respectively.



The image displays two side-by-side mobile app screens for the Malaria Outbreak Early Warning System (MOEWS). The left screen is the login page, featuring the title 'Malaria Outbreak Early Warning System (MOEWS)' at the top. Below the title are input fields for 'Email' and 'Password'. There are three blue buttons: 'LOGIN', 'REGISTER', and 'RESET PASSWORD'. Below these buttons is the text 'or sign in using' followed by a Google logo. The right screen is the registration page, with input fields for 'Name', 'Surname', 'Email Address', 'Phone Number', 'Password', and 'Confirm Password'. It features a blue 'REGISTER' button and a 'Back To Login' link. Both screens have a standard Android navigation bar at the bottom with icons for home, back, and search.

Figure 7.5: MOEWS registration and login screen.

All users first register as normal users. Thereafter, the role of users who have been previously identified as IK experts is manually changed by the researcher in the database to an ambassador role. Due to their IK knowledge, the ambassadors have the right to report observed IK indicators and extreme weather conditions. Normal users can only view the reported IK indicators and extreme weather conditions. Upon successful registration, the user can log in and be taken to the main menu, as seen in Figure 7.6.

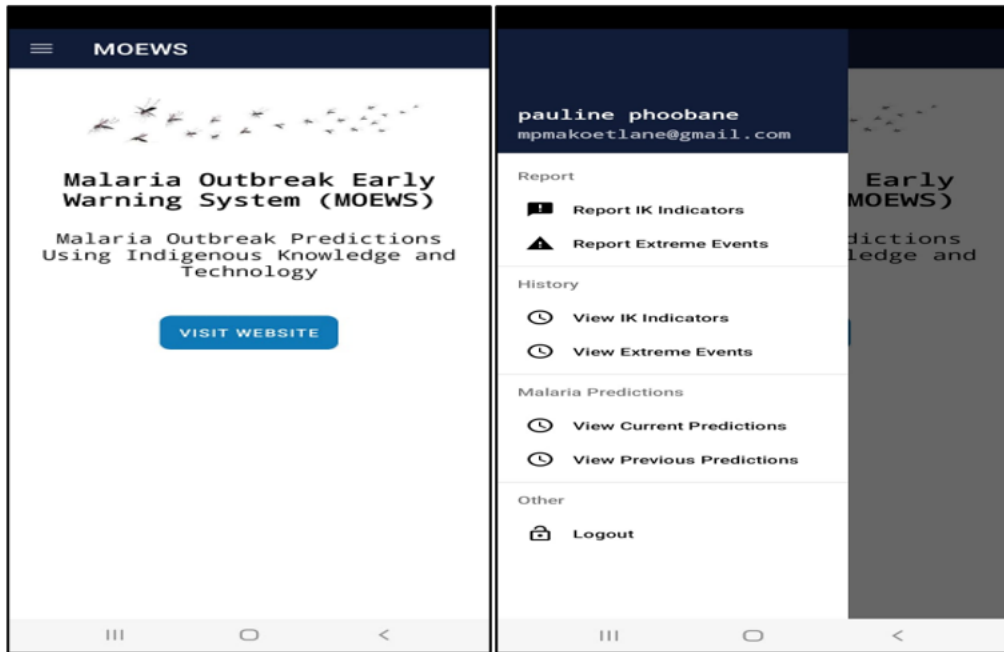


Figure 7.6: MOEWS main menu.

The main menu has a link "visit website" to the system web portal. Upon clicking the 'hamburger' icon on the top left corner of the app, the user is presented with the following options: report IK indicators, report extreme weather, view reported IK indicators, view reported extreme weather, view current malaria predictions, view previous predictions and log out. The subsequent section provides a comprehensive overview of these options.

Report IK indicators sub menu

As previously mentioned, this option is only available to users with an ambassador role. Since the IK indicators are grouped by the four seasons of the year, summer, autumn, winter and spring, the user chooses the season for which to report the observed IK indicators. All known indicators are pre-loaded into the database to ensure consistency and to minimise errors. The ambassador will then be presented with a list of indicators to choose from for the selected season. For each observed IK being reported, the picture of the observed IK is uploaded, the Global Positioning System (GPS) coordinates of the location of the reporting ambassador are also captured and a brief description regarding the IK indicator is provided. After filling in all the information and uploading the picture, the ambassador presses a "SUBMIT" button to store the captured data in the database, as illustrated in Figure 7.7.

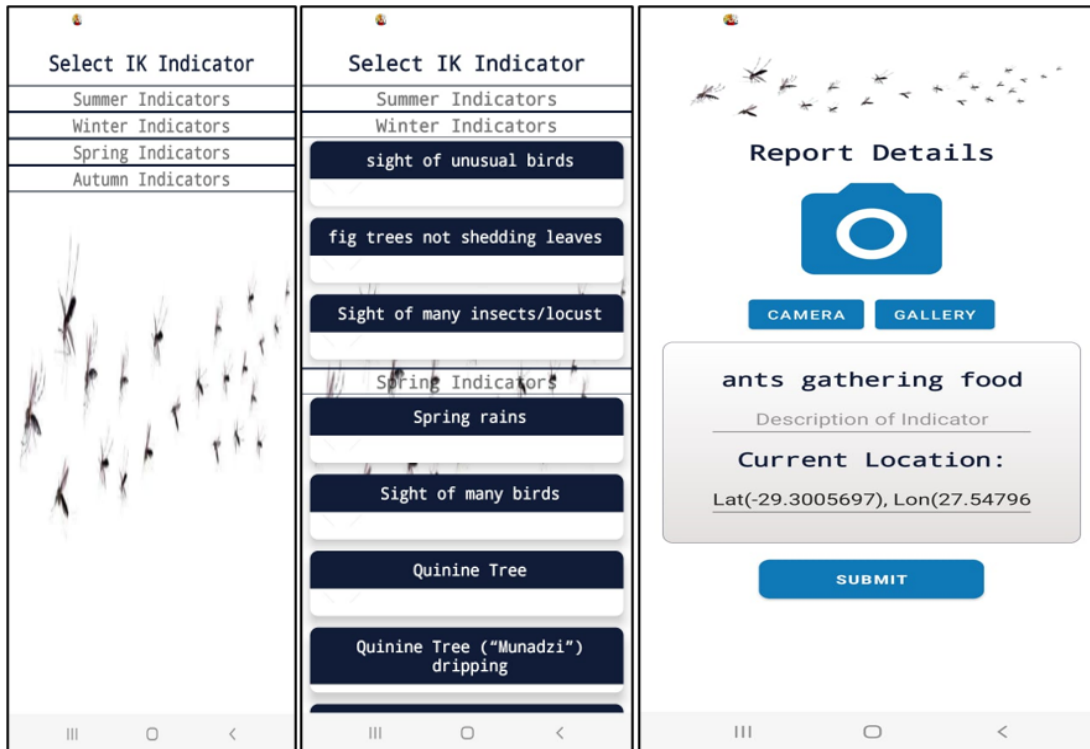
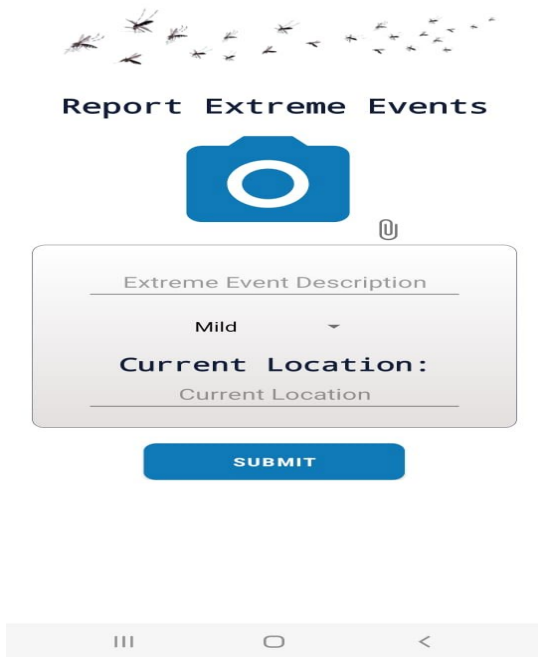


Figure 7.7: MOEWS screens for reporting IK indicators.

The first screen in Figure 7.7 shows categories of the IK indicators by season, while the middle screen shows IK indicators for selected seasons, in this case winter and spring. The last screen on the right shows the interface for capturing the observed IK indicators.

Report Extreme weather sub menu

In a manner similar to reporting IK indicators, an ambassador can also report extreme weather conditions. The event has options, mild, moderate and severe, to express the severity of the observed weather condition. Similarly, the picture of the extreme weather observed and the reporter's location are captured as illustrated in Figure 7.8. The captured data is stored in the database after clicking the 'submit' button.



Report Extreme Events

Extreme Event Description

Mild

Current Location:

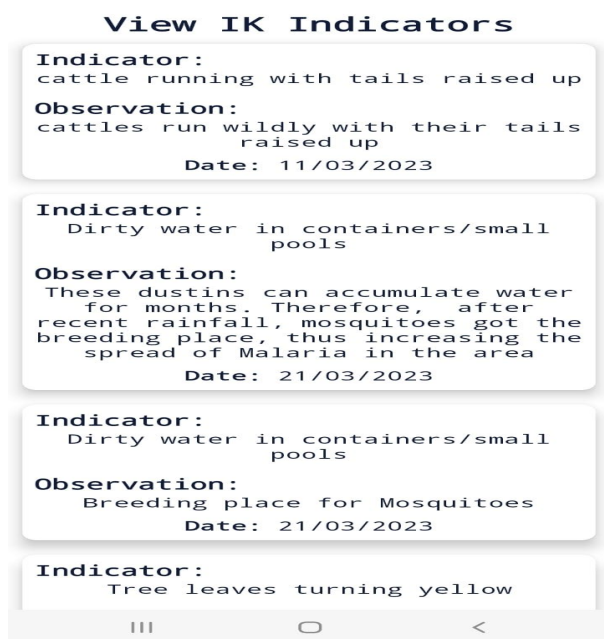
Current Location

SUBMIT

Figure 7.8: MOEWS' report extreme weather screen.

View IK indicators sub menu

This menu opens a screen below to allow registered users to view the observed IK malaria indicators, as reported by the ambassadors. Figure 7.9 below shows some of the reported IK indicators.



View IK Indicators

Indicator:
cattle running with tails raised up

Observation:
cattles run wildly with their tails raised up

Date: 11/03/2023

Indicator:
Dirty water in containers/small pools

Observation:
These dustins can accumulate water for months. Therefore, after recent rainfall, mosquitoes got the breeding place, thus increasing the spread of Malaria in the area

Date: 21/03/2023

Indicator:
Dirty water in containers/small pools

Observation:
Breeding place for Mosquitoes

Date: 21/03/2023

Indicator:
Tree leaves turning yellow

Figure 7.9: MOEWS view reported IK indicator screen.

The screen shows the reported IK indicators such as 'dirty water in containers/small pools', a brief description of the indicator and the date of the report.

View extreme weather sub menu

This sub menu opens the screen showing the reported extreme weather conditions, as depicted in Figure 7.10. The information extracted from the database regarding the reported extreme conditions includes the event, the severity and the date of the report.



Figure 7.10: MOEWS view extreme weather events screen.

View prediction sub menu

This sub menu opens the screens with malaria outbreak predictions for the current month. The forecasts are given from the indigenous knowledge perspective and the modern scientific angle using weather parameters and hydrological drought indices. The IK prediction is seasonal. For example, the screen below shows the autumn malaria outbreak forecast using IK in 2023. Malaria outbreak prediction using weather parameters provides monthly forecasts, which are done one month ahead to account for the time lags between weather parameters and malaria outbreaks. Figures 7.11 and 7.12 show the previous prediction using weather parameters and IK, respectively.

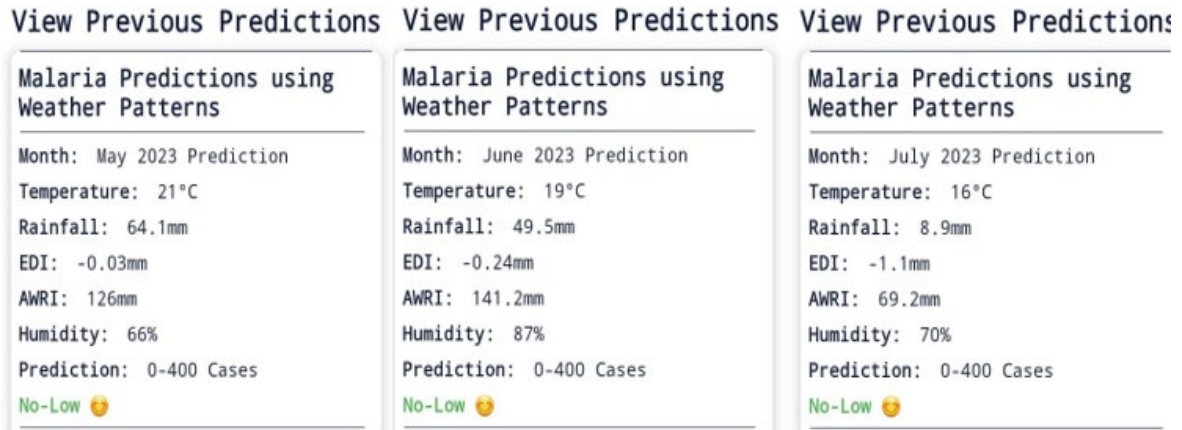


Figure 7.11: MOEWS screens showing the previous predictions done in 2023 using weather patterns.

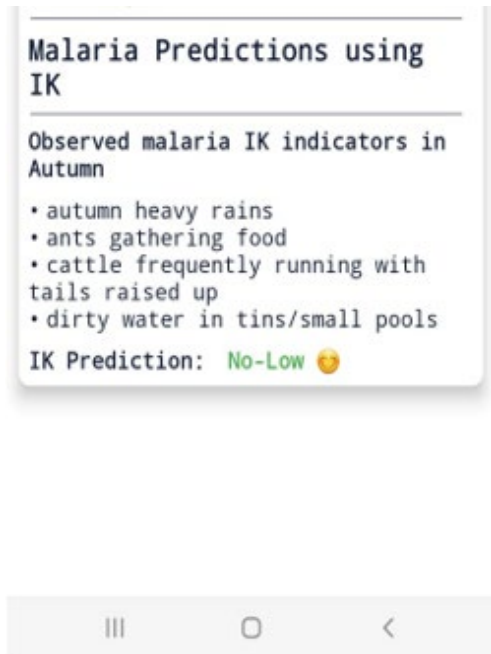


Figure 7.12: MOEWS screens showing the previous predictions done in Autumn 2023 using IK.

The IK prediction was conducted for autumn, whereas the modern science prediction was carried out for three consecutive months: May, June, and July. Each monthly prediction was performed independently. Figure 7.13 demonstrates the prediction for January 2024 carried out in December 2023 using AWRI, EDI and temperature reported in December and rainfall and relative humidity reported in November 2023.

View Current Predictions

Malaria Predictions using Weather Patterns

Month: January 2024 Prediction

Temperature: 19.00°C

Rainfall: 78.70mm

EDI: 0.5mm

AWRI: 134.7mm

Humidity: 61.00%

Prediction: 0-400 Cases

No-Low 🍌

Figure 7. 13: Malaria prediction for January 2024.

7.4 Web Portal

The dissemination and communication of malaria outbreak alerts are also made available to end users on the web portal. The portal also provides comprehensive information about the entire project. The web portal can be accessed at the following link: <http://marksz-001-site1.htempurl.com/>. The portal grants access to both the predictions and the observed parameters utilised in making these predictions, which are keyed by ambassadors via the mobile application. These parameters include IK indicators and meteorological variables such as rainfall, temperature, relative humidity, the Effective Drought Index (EDI) and the Available Water Resource Index (ARWI). This information can be viewed in either text or graphical format, based on the user's preference.

7.4.1 Text View of Malaria Predictions

The same textual prediction information depicted under view prediction (current and previous) was made available on the web portal too. Figure 7.14 depicts the textual view of malaria predictions and input features.

View Current Predictions

Malaria Predictions using Weather Patterns

Month: January 2024 Prediction

Temperature: 19.00°C

Rainfall: 78.70mm

EDI: 0.5mm

AWRI: 134.7mm

Humidity: 61.00%

Prediction: 0-400 Cases

No-Low 🟢

Figure 7.14: malaria prediction in January 2024.

Similar to the prediction views available in the mobile application, both past and current malaria predictions can be accessed through the web portal.

7.4.2 Graphical View of Malaria Outbreak Predictions

Users can opt to view malaria outbreak communications graphically. Figure 7.15 shows weather data recorded over eight months in 2023.

RAINFALL

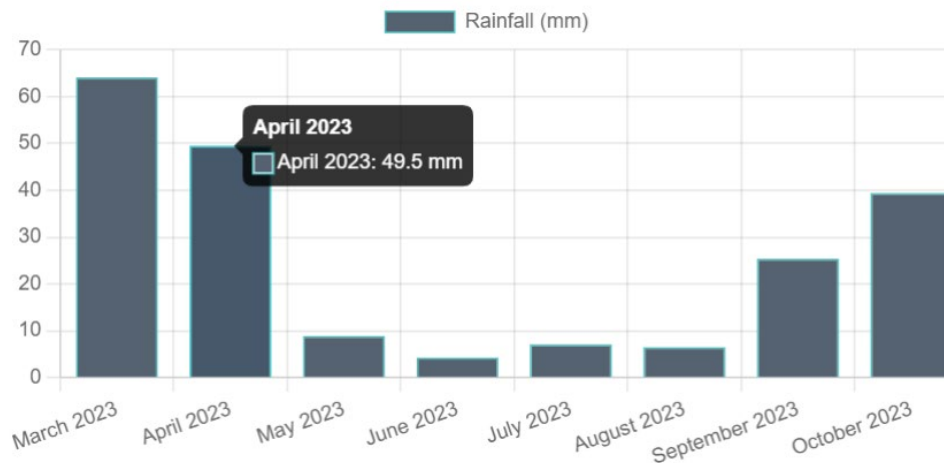


Figure 7.15: graphical view of weather parameters.

By moving the mouse over the graphs, the user can see the actual value of any bar. For example, the value for the bar corresponding to rainfall in April 2023 is 49.5mm. The graphical view shows trends in the weather parameters that contribute to malaria outbreaks. Figure 7.16 displays images of the actual IK indicators observed in Autumn that contributed to malaria predictions based on IK.

OBSERVED INDIGENOUS KNOWLEDGE INDICATORS

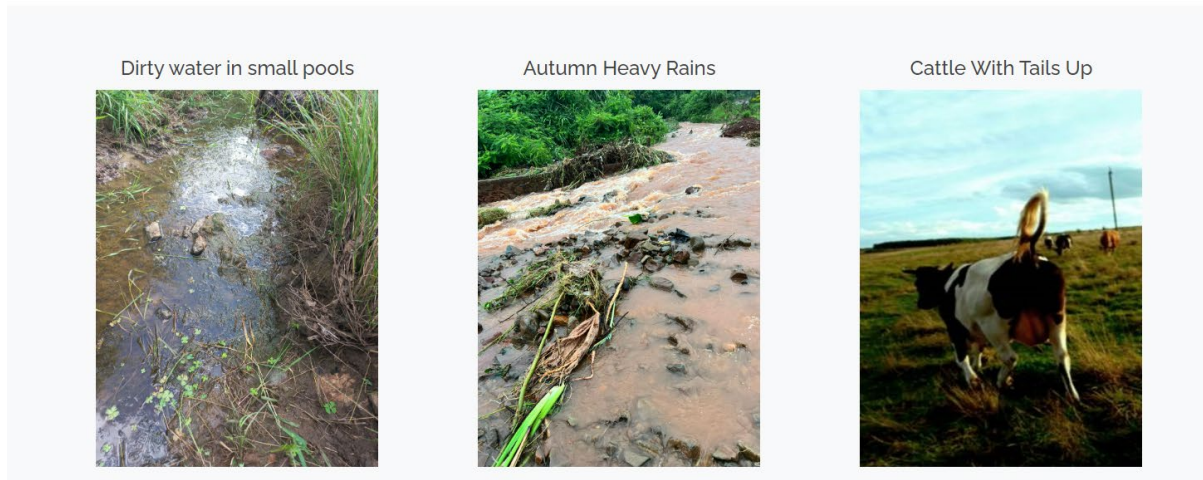


Figure 7.16: IK indicators observed in Autumn 2023.

7.5 Social Media

The researcher leveraged the power of social media to communicate information about malaria. Users following the MOEWS Facebook page can engage with each other in discussions on various malaria-related topics and current information shared on the page. Figure 7.17 shows a MOEWS Facebook page.



Figure 7.17: Facebook page of MOEWS.

A link to the MOEWS web portal is provided for followers on the MOEWS Facebook page, enabling followers to access more information about MOEWS and utilise all its functionalities across various platforms. The researcher is considering exploring other social media channels, such as YouTube, to increase the system's versatility.

7.6 Conclusion

This chapter describes the implementation of MOEWS's third component, which focuses on disseminating and communicating information about malaria outbreaks. It also provides an overview of the MOEWS, which uses modern technology with IK to predict malaria outbreaks.

The use of mobile applications in the communication and dissemination of malaria outbreaks allows greater accessibility, taking advantage of the high levels of mobile ownership by the Vhembe people. The web portal complements this by providing a centralised platform for disseminating alerts and predictions and for displaying observed parameters such as IK indicators and weather data. Users can view this information in textual and graphical formats, enhancing the system's versatility. Moreover, MOEWS also leverages social media platforms, in particular Facebook, to broaden its reach and engaging users in discussions about malaria-related issues.

The next chapter provides the conclusion of this research study and makes recommendations for future studies.

8. Chapter Eight: Conclusion and Future Studies

8.1 Introduction

This final chapter critically examines the research findings, discussing how they align with the study's objectives. The chapter explores the implications of these findings within the broader context of the field, underscoring the novel contributions of this research study. Additionally, the limitations of the study are discussed. Finally, potential avenues for further research are suggested, pointing out ways the findings could be further expanded upon or utilised in future work and policy making.

8.2 Evaluation of the Research Objectives

This sub-section details how the research objectives were achieved.

8.2.1 Climatic Variabilities, IK Indicators and Their Effects on Infectious Disease Outbreaks.

The complex nature of infectious disease prediction necessitates a complete understanding of infectious disease risk knowledge. An investigation was conducted on indigenous knowledge, weather data and historical malaria data. The indigenous knowledge indicators (IKIs) that Vhembe communities rely on to predict malaria outbreaks were investigated. Results indicate that the respondents possessed a rich reservoir of IK for general purposes and also for infectious disease prediction (Phoobane and Masinde, 2023), with over 76% affirming its use for a variety of purposes. Most respondents (75.3%) use IK for farming, with 63.1% using IK for disease prediction. Most participants (74.3%) reported adequate confidence in using IK for general purposes. Respondents who had stayed longer in Vhembe reported higher confidence levels in using IK, as did those above 66 years old in using this knowledge system for disease prediction. The IK indicators for malaria outbreaks, documented in Table 4.13, were stored in a cloud-based database.

Weather data indicate that Vhembe is generally a wet and warm area, as shown by EDI, AWRI, rainfall and temperature, hence exhibiting favourable conditions for mosquitoes to thrive. Through the analysis of mutual information between features, as conducted in this research study, evidence linked these features (independent variables) with malaria outbreaks. However, optimal mutual information was achieved at different time lags for the different variables, as seen in Table 6.4. AWRI, EDI and

temperature showed the highest MI score to the malaria outbreaks at a time lag of one month, while rainfall and relative humidity exhibited its highest MI score at a time lag of two months. AWRI had the highest MI, indicating that it can be used as a measure of the conduciveness of malaria-causing vectors (Phoobane, Masinde and Botai, 2022).

8.2.2 Malaria Early Warning System Framework

The objective of building a Malaria Early Warning System (MOEWS) framework was successfully achieved. The framework was built based on foundational principles from the United Nations International Strategy on Disaster Reduction (UNISDR), adapted to suit the specific requirements of infectious disease outbreaks early warning system with a focus on malaria. The MOEWS framework consists of these three elements: 1. Malaria Risk Knowledge, 2. Malaria Outbreaks Prediction and 3. Malaria Outbreaks Communication and Dissemination.

Data from multiple domains, including conventional weather data, historical malaria data and IK indicators were systematically collected and integrated into the framework. Malaria risk knowledge serves as the foundation for the malaria prediction framework. Including IK indicators in the framework enhanced cultural relevance and ensured community engagement, making the MOEWS scientifically rigorous and socially effective. This augments community understanding and trust in MOEWS and enhances the ability of local communities to respond effectively to malaria outbreaks. The consideration of indigenous coping strategies within MOEWS fosters acceptance of MOEWS among local people, adding a layer to MOEWS' s effectiveness.

Moreover, the MOEWS capitalises on the predictive accuracy of machine learning to predict malaria outbreaks using malaria, weather and IK data. The framework emphasises the importance of timely collection of data in predictive models and communication and dissemination of malaria outbreak alerts. Timely collection of data was accomplished using a mobile application that facilitated the real-time capture of observed IK indicators and dissemination of malaria outbreak alerts to communities and stakeholders. Mobile technology ensured that even remote areas could access vital alerts, allowing for prompt preventive actions.

8.2.3 Fuzzy Cognitive Maps for Malaria Outbreak Prediction

This research study successfully built a fuzzy cognitive map (FCM) to forecast malaria outbreaks using IKIs (Phoobane and Masinde, 2023). The IK indicators that Vhembe participants use to forecast malaria outbreaks were collected and stored in a database. These IK indicators were processed and used to develop FCM models for the four seasons of the year. Firstly, the IKI concepts were determined with each IKI representing a concept in the FCM. Secondly, the causal effects of IKIs among themselves and with malaria were determined to establish the interrelations between IKIs and malaria outbreaks in Vhembe. To achieve this, participants individually rated their perceptions of the causal effects of IKIs on malaria outbreaks, using the following linguistic variables: strong positive, positive, none, negative, or strong negative across the four distinct seasons (summer, autumn, winter, and spring). The responses of the participants on the causal effects of the concepts on other concepts and on malaria outcomes were quantified (see Tables 5.2 through Table 5.5), as the percentage of responses at each category of causal effect (strong positive, positive, none, negative, and strong negative). Mean values derived from participant responses were used to aggregate the perceptions for each concept. Thirdly, we built weighted adjacency matrices, assigning numeric values to the relationships between concepts and with malaria based on their degree of causality as specified by the participants, using linguistic variables. These values, ranging from -1 to 1, represent the strength and direction of influence between concepts, forming a matrix that quantifies these interrelations. Lastly, we used the weighted adjacency matrices to create FCMs for the four seasons of the year.

The importance of each IK indicator within each seasonal FCM was established by analysing the density, in-degree, out-degree and centrality measures. Table 8.1 below shows IK indicators and concepts that play a critical role in predicting malaria outbreaks.

Table 8.1: IK Indicators critical in malaria outbreak prediction in Vhembe.

Season	Concepts	Concept type	In-degree	Out-degree	Centrality
Summer	Summer floods	driver	0	1.6	1.6
	Summer heavy rainfalls	ordinary	6.1	1.5	7.61
	Dirty water in containers/small pools	ordinary	1.599	0.8	2.4

	Summer high temperature	driver	0	1	1
	Summer moon surrounded by clouds	driver	0	1.5	1.5
	Summer sun surrounded by clouds	driver	0	1.3	1.3
Autumn	Autumn heavy rainfalls	ordinary	0.6	1.5	2.1
	Dirty water in the containers/small pools	ordinary	0.8	0.8	1.6
Winter	Fig "Muhuyu" tress not shedding leaves	driver	0	0.6	0.6
	Sight of many insects/locusts	driver	0	0.6	0.6
Spring	Spring rains	ordinary	1.2	0.6	1.8
	"Mofafa" grass having many ticks	driver	0	1	1

The findings highlight that in Autumn, "autumn heavy rainfalls" and "dirty water in containers/small pools" are major predictors of malaria outbreak while "summer heavy rainfalls," "dirty water in containers" and "summer temperature" are key indicators in summer. On the other hand, the main predictors of malaria in winter are 'Fig "Muhuyu" trees not shedding leaves' and 'sight of insects/locusts' while spring rainfalls and "'Mofafa grass" having many ticks' were found important for malaria prediction in spring.

To make predictions using the created FCM models, observed IKIs were collected close to real-time. An Android mobile application was developed and tested to facilitate reporting observed IK indicators. The application is designed to work even in places with no or limited Internet connection. To minimise errors in reporting observed IK indicators, the application allowed 'IK experts' to report observed IK indicators by choosing from a list from the database. Reported IK indicators were extracted from the database and used in a fuzzy cognitive model to predict malaria outbreaks.

8.2.4 Machine Learning Model for Malaria Outbreak Prediction Using Weather Data

The objective of using weather data and drought indices for malaria prediction was achieved through a systematic approach that integrated these environmental factors. The machine learning models were trained and tested on a portion of the data. The study focused on precipitation, temperature and relative humidity. The hydrological drought indices, EDI and AWRI, were also considered. These variables were selected

due to their significant impact on mosquito life cycles and malaria transmission dynamics, making them essential for accurate malaria outbreak predictions.

Historical weather data on Vhembe, spanning 21 years, was provided by SAWS. This dataset provided a solid foundation for analysis, with weather data sourced from specific climatic districts and stations. The computation and use of drought indices from daily precipitation data provided a new perspective on malaria prediction.

It was critical to pre-process the data to allow data integration from heterogeneous sources and prepare data for machine learning, ensuring that the models could effectively learn from the data and make accurate predictions. Data pre-processing included aligning time series data, interpolation of missing values, normalisation and dealing with outliers.

Feature selection and analysis played a vital role in identifying the most relevant variables for prediction. Through mutual information analysis, the research study found that AWRI, EDI and temperature had the strongest associations with malaria incidence. This validation of key variables was essential in enhancing the predictive accuracy of the model.

Various machine learning algorithms involving regression and classification were tested in the development and evaluation phases of building a model. While regression models did not perform as expected, classification models such as MLP, SVM and KNN demonstrated high accuracy, precision, recall and F1 scores. MLP was considered for deployment because it outperformed all the other classification models, achieving an F1 score (a balanced measure of precision and recall) of 96%, precision of 95%, recall of 100% and accuracy of 93%. Figure 8.1 below displays the performances achieved by MLP.

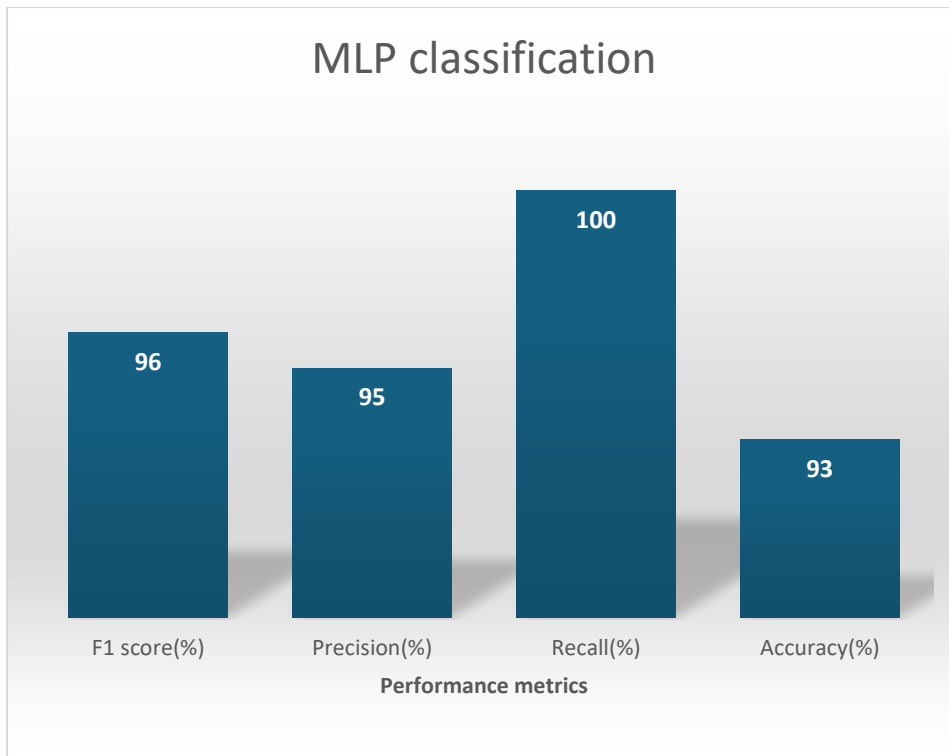


Figure 8. 1: MLP performance during the test phase.

MLP's high performance ensured that the selected weather data and drought indices were effectively used to predict malaria outbreaks.

8.2.4 Building, Testing and Evaluating the System Prototype

We successfully implemented an Android mobile application prototype. The prototype takes care of almost all functionalities of MOEWS. One of the main functionalities of the prototype is to remotely report the observed IKIs. The application is designed to work in areas with no or limited internet. It allows collecting data (observed indigenous knowledge indicators) required to predict malaria outbreaks. The application is synchronised with a cloud-based database to store the reported IKIs automatically. Moreover, the prototype also allows for the reporting of extreme events. The mobile application facilitates the dissemination of malaria outbreak alerts by providing current and previous predictions. A web portal for malaria outbreak communication and dissemination complements the prototype. Figure 8.2 shows the interface of the mobile application prototype, which reflects the application's functionalities.

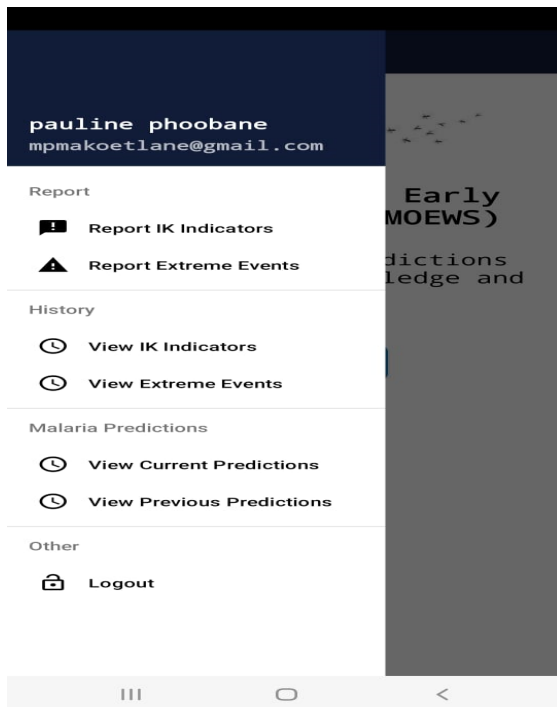


Figure 8. 2: MOEWS mobile application interface showing functionalities of the prototype.

The building of the MOEWS prototype was followed by a testing and evaluation phase to assess its functionality and performance. Firstly, the prototype was tested by the team of developers and researchers at CUT. Secondly, the prototype was given to the Vhembe community to collect data in the autumn of 2023. This allowed us to identify features that were not user-friendly, with adjustments being made accordingly on the prototype.

Using the reported IKIs in autumn 2023 and collected weather parameters from online sources, the predictive element of the model was evaluated. The predictions made by both the IKI component and the component that uses weather data were compared with the malaria incidences reported by the Department of Health in Limpopo during the same evaluation period. Malaria prediction using weather data was done from April 2023 to December 2023. Figure 8.2 below shows the predicted values of the systems against the actual malaria values using MLP classification models. The predicted values are classes expressing malaria categories or severity, as discussed in Chapter 6.

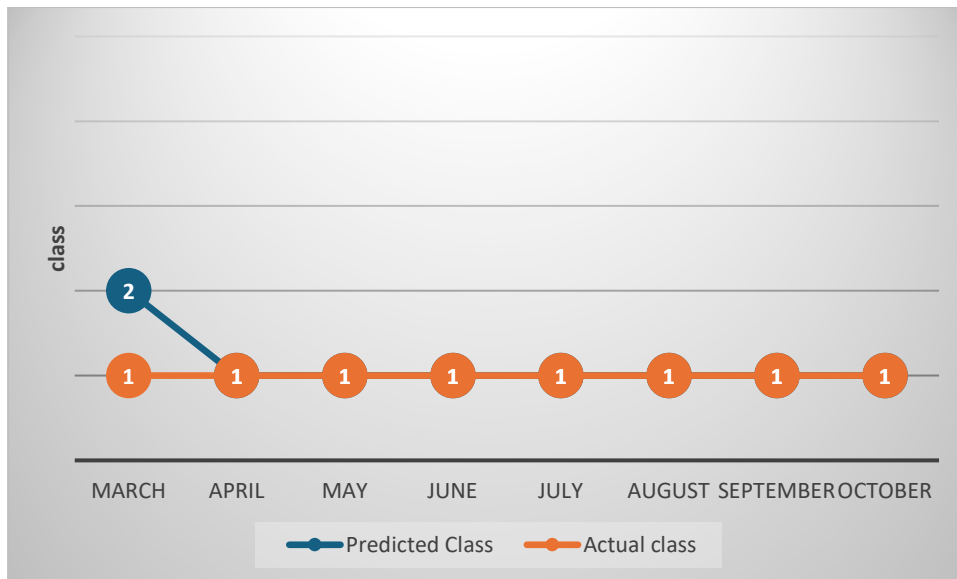


Figure 8. 3: Predicted classes vs actual classes.

8.3 Innovative Contributions of the Research Thesis

This research introduces a ground-breaking approach to enhancing EWS for malaria outbreaks by integrating indigenous knowledge systems (IKS) with advanced ML techniques and weather parameters. The study addresses critical gaps in existing EWS that tend to ignore the community the EWS intend to serve and their existing coping mechanisms. MOEWS offers new insights into disease surveillance, particularly in the context of SSA. This hybrid approach leverages the strengths of modern science and IK, delivering an intelligent, effective and contextually relevant tool tailored to the needs of local African communities.

A significant innovation in this research is the application of FCMs to represent and analyse the causal relationships between malaria indicators derived from indigenous knowledge and actual malaria outbreaks. This method offers an understanding of the factors influencing malaria transmission, allowing for the incorporation of community-based insights into the predictive model. By involving local communities in the development process, the research ensures that the early warning system is not only scientifically robust but also culturally sensitive and accessible to those it is intended to serve.

The research also breaks new ground by incorporating the hydrological drought indices, EDI and the AWRI, into the malaria prediction model. This innovative approach acknowledges the role of water availability in the soil in the lifecycle of malaria vectors,

offering a more reliable prediction model. By mapping these indices to the ability of malaria-carrying vectors to thrive, the research enhances the accuracy of malaria outbreak predictions.

Additionally, the research emphasises a bottom-up approach, integrating community knowledge and coping mechanisms into the early warning system. This focus on local engagement ensures that the system is effective in predicting outbreaks and relevant and empowering for the communities it serves. The MOEWS contributes to a more resilient and sustainable approach to disease surveillance, enabling proactive measures and informed decision-making at the local level.

These innovative contributions collectively advance the field of infectious disease prediction, offering a model that can be adapted and applied to other weather-sensitive diseases in various regions. The integration of diverse knowledge systems, coupled with cutting-edge technological tools, positions this research as a significant step forward in the fight against malaria and other infectious diseases sensitive to climate change.

8.4 Recommendations and Future Directions

8.4.1 Enhancing MOEWS as a framework

Future research could enhance MOEWS by combining outcomes from machine learning, big climate data, and IK into a unified result rather than presenting separate outputs. The system could be designed to function as a 'black box' for the target audience, delivering a single aggregated outcome while retaining the ability to reveal detailed component results and the integration process for users interested in understanding the underlying processes.

8.4.2 Malaria Outbreak Risk Knowledge

First, there is a potential for enhancing data integration. Future work could explore incorporating additional heterogeneous data sources, such as satellite imagery, vegetation indices or social media data, to enrich the predictive capabilities of the early warning system. By broadening the data pool, the system may be able to provide more robust and accurate predictions of malaria outbreaks.

8.4.3 Malaria Prediction

Future studies can explore training the models on datasets from various demographic regions and climate-sensitive infectious diseases to assess their effectiveness further. This approach will help determine the models' scalability and adaptability to different parts of South Africa, Southern Africa and the broader African continent. Expanding the training will provide a more comprehensive evaluation of the potential of the model to generalise across diverse geographic and epidemiological contexts. Additionally, more attention should be given to regression models to improve the accuracy of their predictions. This could also include exploring other regression models. A larger sample could also help regression models to learn patterns between weather parameters and malaria incidences.

8.4.4 Malaria Dissemination and Communication

This component of MOEWS needs innovative enhancements to make the system more accessible. One area for improvement is translating alerts into local languages, ensuring that the information is accessible and understandable to all community members, regardless of their primary language. Additionally, incorporating voice-to-text and text-to-voice functionalities could significantly benefit users with varying levels of literacy or disability. Voice-to-text would allow users to report malaria indicators using spoken language. At the same time, text-to-voice would enable the system to deliver alerts audibly, making the information more accessible to those who may not be able to read or prefer auditory communication.

Although findings from this research study highlighted the high penetration of mobile phones and smartphone ownership in Vhembe, further advancements could be considered. These may include developing SMS-based alert systems and integrating the platform with local radio broadcasts to increase system accessibility. These enhancements would ensure that critical warnings reach all at-risk populations, regardless of their access to digital technology.

List of References

- Abdullah, D.M. and Abdulazeez, A.M. (2021) 'Machine Learning Applications based on SVM Classification A Review', *Qubahan Academic Journal*, 1(2), pp. 81–90. Available at: <https://doi.org/10.48161/qaj.v1n2a50>.
- Abera, T. (2020) 'Indigenous knowledge for sustainability education in Ethiopia', *International Journal of Research in Humanities, Arts and Literature*, 8(1), pp. 39–48.
- Abera, T. and Science, A.A. (2020) 'Indigenous knowledge for sustainability education in ethiopia', (January).
- Absar, N. *et al.* (2022) 'The efficacy of deep learning based LSTM model in forecasting the outbreak of contagious diseases', *Infectious Disease Modelling*, 7(1), pp. 170–183.
- Adamu, Y.A. (2021) 'Malaria Prediction Model Using Machine Learning Algorithms', *Turkish Journal of Computer and Mathematics Education (TURCOMAT)*, 12(10), pp. 488–7496.
- Agusto, F.B., Gumel, A.B. and Parham, P.E. (2015) 'Qualitative assessment of the role of temperature variations on malaria transmission dynamics', *Journal of Biological Systems*, 23(4), pp. 597–630. Available at: <https://doi.org/10.1142/S0218339015500308>.
- Ahmed, M. *et al.* (2017) 'A Survey of Evolution in Predictive Models and Impacting Factors in Customer Churn', *Advances in Data Science and Adaptive Analysis*, 09(03), p. 1750007. Available at: <https://doi.org/10.1142/s2424922x17500073>.
- Akanbi, A. (2019) *Development of Semantics-Based Distributed Middleware for Heterogeneous Data Integration and its Application for Drought*. Central University of Technology, Free State.
- Akanbi, A.K. and Masinde, M. (2018) 'Towards the Development of a Rule-Based Drought Early Warning Expert Systems Using Indigenous Knowledge', *2018 International Conference on Advances in Big Data, Computing and Data Communication Systems, icABCD 2018* [Preprint]. Available at: <https://doi.org/10.1109/ICABCD.2018.8465465>.
- Akoglu, H. (2018) 'User's guide to correlation coefficients', *Turkish Journal of Emergency Medicine*, 18(3), pp. 91–93. Available at: <https://doi.org/10.1016/j.tjem.2018.08.001>.
- Alagoz, O. *et al.* (2010) 'Markov Decision Processes: A Tool for Sequential Decision Making under Uncertainty', *Medical Decision Making*, 30(4), pp. 474–483. Available at: <https://doi.org/10.1177/0272989X09353194>.
- Alagoz, O. (2011) 'Markov Decision Processes: A Tool for Sequential Decision Making under Uncertainty', 30(4), pp. 474–483.

- Alibaba (2021) *Epidemic Prediction Technology*. Available at: <https://www.alibabacloud.com/solutions/health-epidemic-prediction> (Accessed: 7 March 2024).
- Alqaissi, E.Y., Alotaibi, F.S. and Ramzan, M.S. (2022) 'Modern Machine-Learning Predictive Models for Diagnosing Infectious Diseases', *Computational and Mathematical Methods in Medicine*, 2022, pp. 1–13. Available at: <https://doi.org/10.1155/2022/6902321>.
- Altizer, S. *et al.* (2013) 'Climate change and infectious diseases: From evidence to a predictive framework', *Science*, 341(6145), pp. 514–519. Available at: <https://doi.org/10.1126/science.1239401>.
- Amuakwa-Mensah, F., Marbuah, G. and Mubanga, M. (2017) 'Climate variability and infectious diseases nexus: Evidence from Sweden', *Infectious Disease Modelling*, 2(2), pp. 203–217. Available at: <https://doi.org/10.1016/j.idm.2017.03.003>.
- Apostolopoulos, I.D. and Groumpos, P.P. (2023) 'Fuzzy Cognitive Maps: Their Role in Explainable Artificial Intelligence', *Applied Sciences*, 13(6), p. 3412. Available at: <https://doi.org/10.3390/app13063412>.
- Appel, M. *et al.* (2018) 'Open and scalable analytics of large Earth observation datasets: From scenes to multidimensional arrays using SciDB and GDAL', *ISPRS Journal of Photogrammetry and Remote Sensing*, 138, pp. 47–56. Available at: <https://doi.org/10.1016/j.isprsjprs.2018.01.014>.
- Arafiyah, R. and Hermin, F. (2018) 'Data mining for dengue hemorrhagic fever (DHF) prediction with naive Bayes method', *Journal of Physics: Conference Series*, 948(1). Available at: <https://doi.org/10.1088/1742-6596/948/1/012077>.
- Armatas, C.A. *et al.* (2016) 'Opportunities to utilize traditional phenological knowledge to support adaptive management of social-ecological systems vulnerable to changes in climate and fire regimes', *Ecology and Society*, 21(1). Available at: <https://doi.org/10.5751/ES-07905-210116>.
- Asenahabi, B.M. (2019) 'Qualitative research, Mixed method research', *International Journal of Contemporary Applied Researches*, 6(5), pp. 76–87. Available at: www.ijcar.net.
- Assan, J.K. *et al.* (2009) 'Health inequality in resource poor environments and the pursuit of the MDGs: Traditional versus modern healthcare in rural Indonesia', *Journal of Health Management*, 11(1), pp. 93–108. Available at: <https://doi.org/10.1177/097206340901100107>.
- Atteya, W.A., Dahal, Keshav. and Hossain, M.Alamgir. (2010) 'Multi-agent system for early prediction of urinary bladder inflammation disease.', in *In 2010 10th International Conference on Intelligent Systems Design and Applications*. IEEE, pp. 539–544.

- Aydin, B., Akkineni, V. and Angryk, R.A. (2016) 'Modeling and indexing spatiotemporal trajectory data in non-relational databases', *Managing Big Data in Cloud Computing Environments*, pp. 133–162. Available at: <https://doi.org/10.4018/978-1-4666-9834-5.ch006>.
- Babukarthik, R.G. *et al.* (2020) 'Prediction of COVID-19 Using Genetic Deep Learning Convolutional Neural Network (GDCNN)', *IEEE Access*, 8, pp. 177647–177666. Available at: <https://doi.org/10.1109/ACCESS.2020.3025164>.
- Bahl, A. *et al.* (2019) 'Recursive feature elimination in random forest classification supports nanomaterial grouping', *NanoImpact*, 15. Available at: <https://doi.org/10.1016/j.impact.2019.100179>.
- Bansal, S. *et al.* (2016) 'Big data for infectious disease surveillance and modeling', *Journal of Infectious Diseases*, 214(Suppl 4), pp. S375–S379. Available at: <https://doi.org/10.1093/infdis/jiw400>.
- Bassi, A. *et al.* (2020) 'An overview of mobile applications (apps) to support the coronavirus disease 2019 response in India', *Indian Journal of Medical Research*. Wolters Kluwer Medknow Publications, pp. 468–473. Available at: https://doi.org/10.4103/ijmr.IJMR_1200_20.
- Bell, D.G. *et al.* (2007) 'NASA world wind: Opensource GIS for mission operations', *IEEE Aerospace Conference Proceedings* [Preprint]. Available at: <https://doi.org/10.1109/AERO.2007.352954>.
- Bell, E., Bryman, A. and Harley, B. (2022) *Business research methods*. Oxford university press.
- Berry, D.C. (1994) 'Involving users in expert system development', *Expert Systems*, 11(1), pp. 23–28. Available at: <https://doi.org/10.1111/j.1468-0394.1994.tb00314.x>.
- Bhat, P., Hegde, P. and Malaganve, P. (2018) 'A Novel Framework for Big Data Analytics in Business Intelligence', *International Journal of Computer Sciences and Engineering*, 6(12), pp. 855–859. Available at: <https://doi.org/10.26438/ijcse/v6i12.855859>.
- Bukhari, S.N.H., Webber, J. and Mehbodniya, A. (2022) 'Decision tree based ensemble machine learning model for the prediction of Zika virus T-cell epitopes as potential vaccine candidates', *Scientific Reports*, 12(1). Available at: <https://doi.org/10.1038/s41598-022-11731-6>.
- Burgess, R.G. (ed.) (2005) *The ethics of educational research*. Routledge: Palmer Press.
- Byun, H.-R. and Wilhite, D.A. (1999) 'Objective quantification of drought severity and duration.', *Journal of Climate*, 12(9), pp. 2747–2756.
- Cao, L.J. *et al.* (2003) 'A comparison of PCA, KPCA and ICA for dimensionality reduction in support vector machine', *Neurocomputing*, 55(1–2), pp. 321–336. Available at: [https://doi.org/10.1016/S0925-2312\(03\)00433-8](https://doi.org/10.1016/S0925-2312(03)00433-8).

- Caruso, C. (1998) *Interpolation Methods Comparison, Computers Math. Applic.*
- Casey, D. (2006) 'Choosing an appropriate method of data collection. ', *Nurse Researcher*, 13(3).
- Ceccato, P. *et al.* (2018a) 'Data and tools to integrate climate and environmental information into public health', *Infectious Diseases of Poverty*, 7(1), p. 126. Available at: <https://doi.org/10.1186/s40249-018-0501-9>.
- Ceccato, P. *et al.* (2018b) 'Data and tools to integrate climate and environmental information into public health', *Infectious Diseases of Poverty*, 7(1), pp. 1–11. Available at: <https://doi.org/10.1186/s40249-018-0501-9>.
- Cella, W. *et al.* (2019) 'Do climate changes alter the distribution and transmission of malaria? Evidence assessment and recommendations for future studies', *Revista da Sociedade Brasileira de Medicina Tropical*, 52(November), pp. 0–2. Available at: <https://doi.org/10.1590/0037-8682-0308-2019>.
- Chae, S., Kwon, S. and Lee, D. (2018) 'Predicting infectious disease using deep learning and big data', *International Journal of Environmental Research and Public Health*, 15(8). Available at: <https://doi.org/10.3390/ijerph15081596>.
- Chang, K. (2019) 'Geographic Information System', in *International Encyclopedia of Geography*. Wiley, pp. 1–10. Available at: <https://doi.org/10.1002/9781118786352.wbieg0152.pub2>.
- Chen, C., Dubin, R. and Kim, M.C. (2014) 'Emerging trends and new developments in regenerative medicine: A scientometric update (2000-2014)', *Expert Opinion on Biological Therapy*, 14(9), pp. 1295–1317. Available at: <https://doi.org/10.1517/14712598.2014.920813>.
- Chen, M., Mao, S. and Liu, Y. (2014) 'Big data: A survey', *Mobile Networks and Applications*, 19(2), pp. 171–209. Available at: <https://doi.org/10.1007/s11036-013-0489-0>.
- Chen, S. and Bien, J. (2020) 'Valid Inference Corrected for Outlier Removal', *Journal of Computational and Graphical Statistics*, 29(2), pp. 323–334. Available at: <https://doi.org/10.1080/10618600.2019.1660180>.
- Chen, X. *et al.* (2019) 'Discovering thematic change and evolution of utilizing social media for healthcare research', *BMC Medical Informatics and Decision Making*, 19(Suppl 2). Available at: <https://doi.org/10.1186/s12911-019-0757-4>.
- Chicco, D., Warrens, M.J. and Jurman, G. (2021) 'The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation', *PeerJ Computer Science*, 7, p. e623. Available at: <https://doi.org/10.7717/peerj-cs.623>.

- Chien, L.-C. *et al.* (2018) 'Surveillance on the endemic of Zika virus infection by meteorological factors in Colombia: a population-based spatial and temporal study', *BMC Infectious Diseases*, 18(1), p. 180. Available at: <https://doi.org/10.1186/s12879-018-3085-x>.
- Ciritoglu, H.E. *et al.* (2018) 'Towards a Better Replica Management for Hadoop Distributed File System', *Proceedings - 2018 IEEE International Congress on Big Data, BigData Congress 2018 - Part of the 2018 IEEE World Congress on Services*, pp. 104–111. Available at: <https://doi.org/10.1109/BigDataCongress.2018.00021>.
- Cobo, M.J. *et al.* (2011) 'An approach for detecting, quantifying, and visualizing the evolution of a research field: A practical application to the Fuzzy Sets Theory field', *Journal of Informetrics*, 5(1), pp. 146–166. Available at: <https://doi.org/10.1016/j.joi.2010.10.002>.
- Coetzer, R.H. and Adeola, A.M. (2020) 'Assessing the Correlation between Malaria Case Mortality Rates and Access to Health Facilities in the Malaria Region of Vhembe District, South Africa', *Journal of Environmental and Public Health*, 2020, pp. 1–15. Available at: <https://doi.org/10.1155/2020/8973739>.
- Conti, M. *et al.* (2018) 'Internet of Things security and forensics: Challenges and opportunities', *Future Generation Computer Systems*, 78, pp. 544–546. Available at: <https://doi.org/10.1016/j.future.2017.07.060>.
- Creswell, J.W. (2014) *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches*. 4th ed. California: SAGE Publications.
- Danesh, F. and Mardani-Nejad, A. (2020) '1.1 A Historical Overview of Bibliometrics.', *Handbook bibliometrics*, p. 7.
- Darkoh, E.L., Larbi, J.A. and Lawer, E.A. (2017) 'A weather-based prediction model of Malaria prevalence in Amenfi West District, Ghana', *Malaria Research and Treatment*, 2017. Available at: <https://doi.org/10.1155/2017/7820454>.
- Dei, G.J., Hall, B.L. and Rosenberg, D.G. (2000) *Indigenous Knowledges in Global Contexts: Multiple Readings of Our World*. University of Toronto Press.
- Deloitte (2019) *Drowning in data, but starving for insights*. Available at: <https://www2.deloitte.com/ro/en/pages/strategy/articles/drowning-in-data-but-starving-for-insights.html> (Accessed: 19 September 2024).
- Deo, R.C. *et al.* (2017) 'Application of effective drought index for quantification of meteorological drought events: a case study in Australia', *Theoretical and Applied Climatology*, 128(1–2), pp. 359–379. Available at: <https://doi.org/10.1007/s00704-015-1706-5>.
- Dieng, S. *et al.* (2015) 'Geo-epidemiology of malaria incidence in the Vhembe District to guide targeted elimination strategies: a local resurgence'. Available at: <https://doi.org/10.21203/rs.3.rs-2052516/v2>.

Dieng, S. *et al.* (2023) 'Geo-epidemiology of malaria incidence in the Vhembe District to guide targeted elimination strategies, South-Africa, 2015–2018: a local resurgence', *Scientific Reports*, 13(1). Available at: <https://doi.org/10.1038/s41598-023-38147-0>.

Diouf, I. *et al.* (2022) 'Impact of future climate change on malaria in West Africa', *Theoretical and Applied Climatology*, 147(3–4), pp. 853–865. Available at: <https://doi.org/10.1007/s00704-021-03807-6>.

Duarte, F. (2024) *Amount of Data Created Daily*. Available at: <https://explodingtopics.com/blog/data-generated-per-day> (Accessed: 19 September 2024).

Durie, M. (2004a) 'Understanding health and illness: Research at the interface between science and indigenous knowledge', *International Journal of Epidemiology*, 33(5), pp. 1138–1143. Available at: <https://doi.org/10.1093/ije/dyh250>.

Durie, M. (2004b) 'Understanding health and illness: Research at the interface between science and indigenous knowledge', *International Journal of Epidemiology*, 33(5), pp. 1138–1143. Available at: <https://doi.org/10.1093/ije/dyh250>.

Ebi, K.L. *et al.* (2021) 'Hot weather and heat extremes: health risks', *The Lancet*, 398(10301), pp. 698–708. Available at: [https://doi.org/10.1016/S0140-6736\(21\)01208-3](https://doi.org/10.1016/S0140-6736(21)01208-3).

van Eck, N.J. and Waltman, L. (2010) 'Software survey: VOSviewer, a computer program for bibliometric mapping', *Scientometrics*, 84(2), pp. 523–538. Available at: <https://doi.org/10.1007/s11192-009-0146-3>.

van Eck, N.J. and Waltman, L. (2017) 'Citation-based clustering of publications using CitNetExplorer and VOSviewer', *Scientometrics*, 111(2), pp. 1053–1070. Available at: <https://doi.org/10.1007/s11192-017-2300-7>.

Emmanuel, T. *et al.* (2021) 'A survey on missing data in machine learning', *Journal of Big Data*, 8(1), p. 140. Available at: <https://doi.org/10.1186/s40537-021-00516-9>.

Engelbrecht, A.P. (2007) *Computational intelligence: an introduction*. 2nd Editio. John Wiley & Sons.

Fang, X. *et al.* (2020) 'Forecasting incidence of infectious diarrhea using random forest in Jiangsu Province, China', *BMC Infectious Diseases*, 20(1). Available at: <https://doi.org/10.1186/s12879-020-4930-2>.

Fatehkia, M., Kashyap, R. and Weber, I. (2018) 'Using Facebook ad data to track the global digital gender gap', *World Development*, 107, pp. 189–209. Available at: <https://doi.org/10.1016/j.worlddev.2018.03.007>.

Felix, G. *et al.* (2019) 'A review on methods and software for fuzzy cognitive maps', *Artificial Intelligence Review*, 52(3), pp. 1707–1737. Available at: <https://doi.org/10.1007/s10462-017-9575-1>.

- Fenollar, F. and Mediannikov, O. (2018) 'Emerging infectious diseases in Africa in the 21st century', *New Microbes and New Infections*, 26, pp. S10–S18. Available at: <https://doi.org/10.1016/j.nmni.2018.09.004>.
- Field, A.P. (2005) *Discovering statistics using SPSS : (and sex, drugs and rock 'n' roll)*. Sage Publications.
- Flick, U. (2015) *Introducing research methodology: A beginner's guide to doing a research project*. SAGE.
- Fu, Y. *et al.* (2023) 'A bibliometric analysis of systematic reviews and meta-analyses in ophthalmology', *Frontiers in Medicine*, 10. Available at: <https://doi.org/10.3389/fmed.2023.1135592>.
- Gaddy, H.G. (2020) 'Using local knowledge in emerging infectious disease research', *Social Science & Medicine*, 258, p. 113107. Available at: <https://doi.org/10.1016/j.socscimed.2020.113107>.
- García, S., Luengo, J. and Herrera, F. (2015) 'Feature selection', *Intelligent Systems Reference Library*, 72(6), pp. 163–193. Available at: https://doi.org/10.1007/978-3-319-10247-4_7.
- García, V. *et al.* (2020) 'Understanding the apparent superiority of over-sampling through an analysis of local information for class-imbalanced data', *Expert Systems with Applications*, 158, p. 113026. Available at: <https://doi.org/10.1016/j.eswa.2019.113026>.
- Garro, A. *et al.* (2018) 'Intelligent agents and environment', *Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics*, 1–3(January), pp. 309–314. Available at: <https://doi.org/10.1016/B978-0-12-809633-8.20327-0>.
- Geneva: World Health Organisation (2020) *World Malaria Report 2020. Years of progress and challenges*. Available at: [https://doi.org/10.1002/\(SICI\)1096-8628\(19971128\)73:1<1::AID-AJMG1>3.0.CO;2-Y](https://doi.org/10.1002/(SICI)1096-8628(19971128)73:1<1::AID-AJMG1>3.0.CO;2-Y).
- GeoPoll (2019) 'Mobile Phone Penetration Throughout Sub-Saharan Africa.' Available at: <https://www.geopoll.com/blog/mobile-phone-penetration-africa/>.
- Githeko, A. *et al.* (2018) 'The launch and operation of the malaria epidemic early warning system in Kenya', *SPG BioMed* [Preprint]. Available at: <https://doi.org/10.32392/biomed.20>.
- Githeko, A.K. *et al.* (2014) 'Development and validation of climate and ecosystem-based early malaria epidemic prediction models in East Africa', *Malaria Journal*, 13(1). Available at: <https://doi.org/10.1186/1475-2875-13-329>.
- Githeko, A.K. and Ndegwa, W. (2001) 'Githeko_and_Ndegwa_Epidemic_prediction2 (1)', 2(1).
- Goddard, W. and Melville, S. (2004) *Research methodology: An introduction*. 2nd edn. Lansdowne: Juta and Company Ltd.

- González, S. *et al.* (2020) 'A practical tutorial on bagging and boosting based ensembles for machine learning: Algorithms, software tools, performance study, practical perspectives and opportunities', *Information Fusion*, 64, pp. 205–237. Available at: <https://doi.org/10.1016/j.inffus.2020.07.007>.
- Goodman, A., Kamath, C. and Kumar, V. (2007) 'Data Analysis in the Twenty-First Century'.
- Goswami, P. *et al.* (2014) 'Relative Roles of Weather Variables and Change in Human Population in Malaria: Comparison over Different States of India', *PLOS ONE*, 9(6). Available at: <https://doi.org/10.1371/journal.pone.0099867>.
- Gray, S. and Cox, L. (2022) *MENTAL MODELER: A tool for environmental planning and research*.
- Grenier, L. (1998) *Working with Indigenous Knowledge: A Guide for Researchers*. Ottawa, Canada: International Development Research Centre.
- Groumpos, P. (2021) 'Modelling COVID-19 using Fuzzy Cognitive Maps (FCM)', *EAI Endorsed Transactions on Bioengineering and Bioinformatics*, 1(2), p. 168728. Available at: <https://doi.org/10.4108/eai.24-2-2021.168728>.
- GSMA (2023) *The Mobile Economy Sub-Saharan Africa 2023*. Available at: www.gsmaintelligence.com.
- GSMA (2024) *The mobile economy*. Available at: <https://www.gsma.com/mobileeconomy/sub-saharan-africa/> (Accessed: 19 September 2022).
- Gu, D. *et al.* (2018) *Deep Learning Project Report*.
- Guo, P. *et al.* (2017) 'Developing a dengue forecast model using machine learning: A case study in China', *PLoS Neglected Tropical Diseases*, 11(10), pp. 1–22. Available at: <https://doi.org/10.1371/journal.pntd.0005973>.
- Gupta, A. *et al.* (2020) 'Heart disease prediction using classification (naive bayes)', in *In Proceedings of First International Conference on Computing, Communications, and Cyber-Security (IC4S 2019)*. Springer Singapore, pp. 561–573.
- Gurcan, F. and Berigel, M. (2018) 'Real-Time Processing of Big Data Streams: Lifecycle, Tools, Tasks, and Challenges', *ISMSIT 2018 - 2nd International Symposium on Multidisciplinary Studies and Innovative Technologies, Proceedings* [Preprint]. Available at: <https://doi.org/10.1109/ISMSIT.2018.8567061>.
- Han, Z., Zhang, Q. and Wen, F. (1999) 'A survey on rough set theory and its application', *Kongzhi Lilun Yu Yingyong/Control Theory and Applications*, 16(2), pp. 323–333. Available at: <https://doi.org/10.1016/j.trit.2016.11.001>.

Harvey, D., Valkenburg, W. and Amara, A. (2021) 'Predicting malaria epidemics in Burkina Faso with machine learning', *PLOS ONE*, 16(6), p. e0253302. Available at: <https://doi.org/10.1371/journal.pone.0253302>.

Hassani, H., Huang, X. and Silva, E. (2019) 'Big Data and Climate Change', *Big Data and Cognitive Computing*, 3(1), p. 12. Available at: <https://doi.org/10.3390/bdcc3010012>.

Health Organization, W. (2022) *World malaria report 2022*. Available at: <https://www.wipo.int/amc/en/>.

Hevner, A. and Chatterjee, S. (2010) *Design Research in Information Systems*. In *Design research in information systems*. Edited by Sharda.R. and Voß.S. And. Boston: Springer.

Hoshen, M.B. and Morse, A.P. (2004) 'A weather-driven model of malaria transmission', *Malaria journal*, 3(1), pp. 1–14.

Hoyos, W., Aguilar, J. and Toro, M. (2022) 'A clinical decision-support system for dengue based on fuzzy cognitive maps', *Health Care Management Science*, 25(4), pp. 666–681. Available at: <https://doi.org/10.1007/s10729-022-09611-6>.

Hu, F. *et al.* (2018) 'ClimateSpark: An in-memory distributed computing framework for big climate data analytics', *Computers and Geosciences*, 115(April), pp. 154–166. Available at: <https://doi.org/10.1016/j.cageo.2018.03.011>.

Hu, Z. *et al.* (2017a) 'A Spatiotemporal Indexing Approach for Efficient Processing of Big Array-based Climate Data with MapReduce', *International Journal of Geographical Information Science*, 31(1), pp. 17–35.

Hu, Z. *et al.* (2017b) 'A Spatiotemporal Indexing Approach for Efficient Processing of Big Array-based Climate Data with MapReduce A spatiotemporal indexing approach for efficient processing of big array-based climate data with MapReduce A Spatiotemporal Indexing Approach for Ef', *International Journal of Geographical Information Science*, 31(1), pp. 17–35.

Hung, T., Nguyen, W.C. and Walker, E.A. (2018) *A First Course in Fuzzy Logic Front Cover*.

Hussain-Alkhateeb, L. *et al.* (2018) 'Early warning and response system (EWARS) for dengue outbreaks: Recent advancements towards widespread applications in critical settings', *PLOS ONE*, 13(5), p. e0196811. Available at: <https://doi.org/10.1371/journal.pone.0196811>.

Hussain-Alkhateeb, L. *et al.* (2021) 'Early warning systems (EWSs) for chikungunya, dengue, malaria, yellow fever, and Zika outbreaks: What is the evidence? A scoping review', *PLOS Neglected Tropical Diseases*, 15(9), p. e0009686. Available at: <https://doi.org/10.1371/journal.pntd.0009686>.

- Immergluck, L.C. *et al.* (2019) 'Geographic surveillance of community associated MRSA infections in children using electronic health record data', *BMC Infectious Diseases*, 19(1), pp. 1–12. Available at: <https://doi.org/10.1186/s12879-019-3682-3>.
- Isaac Abiodun, O. *et al.* (2018) 'State-of-the-art in artificial neural network applications: A survey', *Heliyon*, 4, p. e00938. Available at: <https://doi.org/10.1016/j.heliyon.2018>.
- Jebli, I. *et al.* (2021) 'Prediction of solar energy guided by pearson correlation using machine learning', 224.
- Jewell, C.P. *et al.* (2009) 'A novel approach to real-time risk prediction for emerging infectious diseases: A case study in Avian Influenza H5N1', *Preventive Veterinary Medicine*, 91(1), pp. 19–28. Available at: <https://doi.org/10.1016/j.prevetmed.2009.05.019>.
- Jiang, F. *et al.* (2017) 'Artificial intelligence in healthcare: Past, present and future', *Stroke and Vascular Neurology*, 2(4), pp. 230–243. Available at: <https://doi.org/10.1136/svn-2017-000101>.
- Jiri, O. *et al.* (2016) 'Seasonal Climate Prediction and Adaptation Using Indigenous Knowledge Systems in Agriculture Systems in Southern Africa: A Review', *Journal of Agricultural Science*, 8(5), p. 156. Available at: <https://doi.org/10.5539/jas.v8n5p156>.
- Kahn, G. *et al.* (2017) 'Uncertainty-Aware Reinforcement Learning for Collision Avoidance'.
- Kalipe, G., Gautham, V. and Behera, R.K. (2018) 'Predicting Malarial Outbreak using Machine Learning and Deep Learning Approach: A Review and Analysis', in *2018 International Conference on Information Technology (ICIT)*. IEEE, pp. 33–38. Available at: <https://doi.org/10.1109/ICIT.2018.00019>.
- Khoirunnisa, A. and Ramadhan, N.G. (2023) 'Improving malaria prediction with ensemble learning and robust scaler: An integrated approach for enhanced accuracy', *JURNAL INFOTEL*, 15(4), pp. 326–334. Available at: <https://doi.org/10.20895/infotel.v15i4.1056>.
- Kifle, M.M. *et al.* (2019) 'Malaria Risk Stratification and Modeling the Effect of Rainfall on Malaria Incidence in Eritrea', *Journal of Environmental and Public Health*, 2019. Available at: <https://doi.org/10.1155/2019/7314129>.
- Kijazi A. L.1, Chang'a L. B.1, Liwenga E. T. 2, K.A. 1 and N.S.J. (2013) 'The use of indigenous knowledge in weather and climate prediction in Mahenge and Ismani wards, Tanzania', *Journal of Geography and Regional Planning*, 6(7), pp. 274–279. Available at: <https://doi.org/10.5897/jgrp2013.0386>.
- Kijazi, A.L. *et al.* (2013) 'The use of indigenous knowledge in weather and climate prediction in Mahenge and Ismani wards, Tanzania', *Journal of Geography and Regional Planning*, 6(7), pp. 274–279. Available at: <https://doi.org/10.5897/jgrp2013.0386>.

- Kim, D.W., Byun, H.R. and Choi, K.S. (2009) 'Evaluation, modification, and application of the Effective Drought Index to 200-Year drought climatology of Seoul, Korea', *Journal of Hydrology*, 378(1–2), pp. 1–12. Available at: <https://doi.org/10.1016/j.jhydrol.2009.08.021>.
- Kim, Y. *et al.* (2019) 'Malaria predictions based on seasonal climate forecasts in South Africa: A time series distributed lag nonlinear model', *Scientific Reports*, 9(1), pp. 1–10. Available at: <https://doi.org/10.1038/s41598-019-53838-3>.
- Kolawole, O.D. *et al.* (2016) 'Climate variability and rural livelihoods: How households perceive and adapt to climatic shocks in the Okavango Delta, Botswana', *Weather, Climate, and Society*, 8(2), pp. 131–145. Available at: <https://doi.org/10.1175/WCAS-D-15-0019.1>.
- Kom, Z. *et al.* (2020) 'Determinants of small-scale farmers' choice and adaptive strategies in response to climatic shocks in Vhembe District, South Africa', *GeoJournal*, 87(2), pp. 677–700. Available at: <https://doi.org/10.1007/s10708-020-10272-7>.
- Kosko, B. (1986) 'Fuzzy cognitive maps', *International Journal of Man-Machine Studies*, 24(1), pp. 65–75. Available at: [https://doi.org/10.1016/S0020-7373\(86\)80040-2](https://doi.org/10.1016/S0020-7373(86)80040-2).
- Kothari, C.R. (2004) *Research methodology: Methods and techniques*. New Age International.
- Kothari, R. and Dong, M. (2001) 'Decision trees for classification: A review and some new results. In Pattern recognition: from classical to modern approaches', pp. 169–184.
- Kraemer, M.U.G. *et al.* (2017) 'Spread of yellow fever virus outbreak in Angola and the Democratic Republic of the Congo 2015–16: a modelling study', *The Lancet Infectious Diseases*, 17(3), pp. 330–338. Available at: [https://doi.org/10.1016/S1473-3099\(16\)30513-8](https://doi.org/10.1016/S1473-3099(16)30513-8).
- Krittanawong, C. *et al.* (2017) 'Artificial Intelligence in Precision Cardiovascular Medicine', *Journal of the American College of Cardiology*, 69(21), pp. 2657–2664. Available at: <https://doi.org/10.1016/j.jacc.2017.03.571>.
- Kumar, R. (2019) *Research Methodology a step by step guide for beginners*. fifth edit. SAGE publications.
- Kumar Sharma, S. *et al.* (2021) *An Improved Random Forest Algorithm for Predicting the COVID-19 Pandemic Patient Health*. Available at: <http://annalsofrscb.ro>.
- Kurane, I. (2010) 'The Effect of Global Warming on Infectious Diseases', *Osong Public Health and Research Perspectives*, 1(1), pp. 4–9. Available at: <https://doi.org/10.1016/j.phrp.2010.12.004>.
- Land Jr, W.H. *et al.* (2020) 'The support vector machine.' *The Art and Science of Machine Intelligence: With An Innovative Application for Alzheimer's Detection from Speech* , pp. 45–76.

Leal Filho, W. *et al.* (2022) 'The role of indigenous knowledge in climate change adaptation in Africa', *Environmental Science and Policy*, 136, pp. 250–260. Available at: <https://doi.org/10.1016/j.envsci.2022.06.004>.

Lee, S.W. (2022) 'Regression analysis for continuous independent variables in medical research: statistical standard and guideline of Life Cycle Committee', *Life Cycle*, 2. Available at: <https://doi.org/10.54724/lc.2022.e3>.

Li, W. and Wang, S. (2017) 'PolarGlobe: A web-wide virtual globe system for visualizing multidimensional, time-varying, big climate data', *International Journal of Geographical Information Science*, 31(8), pp. 1562–1582. Available at: <https://doi.org/10.1080/13658816.2017.1306863>.

Li, Z. *et al.* (2015) 'Enabling big geoscience data analytics with a cloud-based, mapreduce-enabled and service-oriented workflow framework', *PLoS ONE*, 10(3), pp. 1–23. Available at: <https://doi.org/10.1371/journal.pone.0116781>.

Li, Z. *et al.* (2017) 'A high performance query analytical framework for supporting data-intensive climate studies', *Computers, Environment and Urban Systems*, 62(2017), pp. 210–221. Available at: <https://doi.org/10.1016/j.compenvurbsys.2016.12.003>.

Li, Z. *et al.* (2019) 'SOVAS: a scalable online visual analytic system for big climate data analysis', *International Journal of Geographical Information Science* [Preprint], (2019). Available at: <https://doi.org/10.1080/13658816.2019.1605073>.

Liang, R. *et al.* (2020) 'Prediction for global African swine fever outbreaks based on a combination of random forest algorithms and meteorological data', *Transboundary and Emerging Diseases*, 67(2), pp. 935–946. Available at: <https://doi.org/10.1111/tbed.13424>.

Liao, Y. *et al.* (2017) 'A new method for assessing the risk of infectious disease outbreak', *Scientific Reports*, 7(January), pp. 1–12. Available at: <https://doi.org/10.1038/srep40084>.

Lio, W. and Liu, B. (2020) 'Uncertain maximum likelihood estimation with application to uncertain regression analysis', *Soft Computing*, 24(13), pp. 9351–9360. Available at: <https://doi.org/10.1007/s00500-020-04951-3>.

Little, R.J. and Rubin, D.B. (2019) *Statistical analysis with missing data*. 3rd edn. New Jersey: John Wiley & Sons.

De Long, D.W. and Fahey, L. (2000) 'Diagnosing cultural barriers to knowledge management', *Academy of Management Perspectives*, 14(4), pp. 113–127. Available at: <https://doi.org/10.5465/ame.2000.3979820>.

Lourenço, J. *et al.* (2018) 'Epidemiology of the Zika virus outbreak in the Cabo Verde islands, West Africa.', *PLoS currents*, 10(2018).

- Lu, E. (2009) 'Determining the start, duration, and strength of flood and drought with daily precipitation: Rationale', *Geophysical Research Letters*, 36(12). Available at: <https://doi.org/10.1029/2009GL038817>.
- Lü, G. *et al.* (2019) 'Reflections and speculations on the progress in Geographic Information Systems (GIS): a geographic perspective', *International Journal of Geographical Information Science*, 33(2), pp. 346–367. Available at: <https://doi.org/10.1080/13658816.2018.1533136>.
- Luther, J. *et al.* (2017) 'World Meteorological Organization (WMO)—Concerted International Efforts for Advancing Multi-hazard Early Warning Systems', in *Advancing Culture of Living with Landslides*. Cham: Springer International Publishing, pp. 129–141. Available at: https://doi.org/10.1007/978-3-319-59469-9_9.
- Macherera, M. and Chimbari, M.J. (2016) 'A review of studies on community based early warning systems', *Jamba: Journal of Disaster Risk Studies*, 8(1), pp. 1–10. Available at: <https://doi.org/10.4102/jamba.v8i1.206>.
- Macherera, M., Chimbari, M.J. and Mukaratirwa, S. (2017) 'Indigenous environmental indicators for malaria: A district study in Zimbabwe', *Acta Tropica*, 175, pp. 50–59. Available at: <https://doi.org/10.1016/j.actatropica.2016.08.021>.
- Mafongoya, P.L. *et al.* (2017) 'Using indigenous knowledge for seasonal quality prediction in managing climate risk in sub-Saharan Africa', in *INDIGENOUS KNOWLEDGE SYSTEMS AND CLIMATE CHANGE MANAGEMENT IN AFRICA*, pp. 43–67.
- Mafongoya, P.L. and Ajayi, O.C. (2017) *Indigenous knowledge systems and climate change management in Africa*, Publications.Cta.Int.
- Mahoo, H. *et al.* (2015) 'Integrating Indigenous Knowledge with Scientific Seasonal Forecasts for Climate Risk Management in Lushoto District in Tanzania', (103).
- Makinde, O.S. and Abiodun, G.J. (2020) 'The impact of rainfall and temperature on malaria dynamics in the KwaZulu-Natal province, South Africa', *Communications in Statistics: Case Studies, Data Analysis and Applications*, 6(2), pp. 97–108. Available at: <https://doi.org/10.1080/23737484.2019.1699000>.
- Malapane, O.L. *et al.* (2022) 'Bibliometric Analysis and Systematic Review of Indigenous Knowledge from a Comparative African Perspective: 1990–2020', *Land*, 11(8). Available at: <https://doi.org/10.3390/land11081167>.
- Malik, A. *et al.* (2021) 'Analysis of dry and wet climate characteristics at Uttarakhand (India) using effective drought index Sustainable Developments in Civil', 105, pp. 1643–1662. Available at: <https://doi.org/10.1007/s11069-020-04370-5>.
- Malik, A. and Kumar, A. (2020a) 'Meteorological drought prediction using heuristic approaches based on effective drought index: a case study in Uttarakhand', *Arabian Journal of Geosciences*, 13(6), p. 276. Available at: <https://doi.org/10.1007/s12517-020-5239-6>.

- Malik, A. and Kumar, A. (2020b) 'Meteorological drought prediction using heuristic approaches based on effective drought index: a case study in Uttarakhand', *Arabian Journal of Geosciences*, 13(6). Available at: <https://doi.org/10.1007/s12517-020-5239-6>.
- Maniruzzaman, Md. *et al.* (2018) 'Accurate Diabetes Risk Stratification Using Machine Learning: Role of Missing Value and Outliers', *Journal of Medical Systems*, 42(5), p. 92. Available at: <https://doi.org/10.1007/s10916-018-0940-7>.
- Manogaran, G. and Lopez, D. (2017) 'Disease surveillance system for big climate data processing and Dengue transmission', *International Journal of Ambient Computing and Intelligence*, 8(2), pp. 88–105. Available at: <https://doi.org/10.4018/IJACI.2017040106>.
- Manogaran, G., Lopez, D. and Chilamkurti, N. (2018) 'In-Mapper combiner based MapReduce algorithm for processing of big climate data', *Future Generation Computer Systems*, 86, pp. 433–445. Available at: <https://doi.org/10.1016/j.future.2018.02.048>.
- Markham, S.K., Kowolenko, M. and Michaelis, T.L. (2015) 'Unstructured Text Analytics to Support New Product Development Decisions.', *Research Technology Management*, 58(2), pp. 30–38.
- Marr, B. (2019) *Artificial Intelligence in Practice: How 50 Successful Companies Used AI and Machine Learning to Solve Problems*. West Sussex: Johns Willey and Sons.
- Martins, R.M. and Von Wangenheim, C.G. (2023) 'Findings on Teaching Machine Learning in High School: A Ten-Year Systematic Literature Review', *Informatics in Education*. Vilnius University Institute of Data Science and Digital Technologies, pp. 421–440. Available at: <https://doi.org/10.15388/infedu.2023.18>.
- Mashego, S.L., Maditsi, M.E. and Bhuda, M.T. (2021) 'A response to Covid-19: Perspectives from African indigenous knowledge systems.', *Gender and Behaviour*, 19(1), pp. 17242–17254.
- Masinde, M. (2012) *Bridge between African Indigenous knowledge and modern science on drought prediction*.
- Masinde, M. (2014) 'An effective drought early warning system for sub-Saharan Africa: Integrating modern and indigenous approaches', *ACM International Conference Proceeding Series*, 28-Septemb(September), pp. 60–69. Available at: <https://doi.org/10.1145/2664591.2664629>.
- Masinde, M. (2020) 'Africa's Malaria Epidemic Predictor: Application of Machine Learning on Malaria Incidence and Climate Data', in *In Proceedings of the 2020 the 4th International Conference on Compute and Data Analysis*, pp. 29–37.
- Masinde, M., Mwagha, M. and Tadesse, T. (2018) 'Downscaling africa's drought forecasts through integration of indigenous and scientific drought forecasts using fuzzy cognitive

maps', *Geosciences (Switzerland)*, 8(4). Available at:
<https://doi.org/10.3390/geosciences8040135>.

De Mauro, A., Greco, M. and Grimaldi, M. (2016) 'A formal definition of Big Data based on its essential features', *Library Review*, 65(3), pp. 122–135. Available at:
<https://doi.org/10.1108/LR-06-2015-0061>.

Mbele, M. and Masinde, M. (2020) 'Mitigating the Impacts of Environmental Pollution in Lejweleputswa District Through Integration of Local and Scientific Knowledge', in, pp. 77–88. Available at: https://doi.org/10.1007/978-3-030-51051-0_6.

Mbewe, M., Phiri, A. and Siyambango, N. (2019) 'Indigenous Knowledge Systems for Local Weather Predictions: A Case of Mukonchi Chieftdom in Zambia', *Environment and Natural Resources Research*, 9(2), p. 16. Available at: <https://doi.org/10.5539/enrr.v9n2p16>.

Menni, C. *et al.* (2020) 'Real-time tracking of self-reported symptoms to predict potential COVID-19', *Nature Medicine*, 26(7), pp. 1037–1040. Available at:
<https://doi.org/10.1038/s41591-020-0916-2>.

Modu, B., Polovina, N., Lan, Y., Konur, S., Asyhari, A.T., *et al.* (2017) 'Towards a Predictive Analytics-Based Intelligent Malaria Outbreak Warning System', *APPLIED SCIENCES-BASEL*, 7(8). Available at: <https://doi.org/10.3390/app7080836>.

Modu, B., Polovina, N., Lan, Y., Konur, S., Taufiq, A.A., *et al.* (2017) 'Towards a predictive analytics-based intelligent malaria outbreakwarning system', *Applied Sciences (Switzerland)*, 7(8), pp. 1–20. Available at: <https://doi.org/10.3390/app7080836>.

Mohapatra, P. *et al.* (2022) 'Determining suitable machine learning classifier technique for prediction of malaria incidents attributed to climate of Odisha', *International Journal of Environmental Health Research*, 32(8), pp. 1716–1732. Available at:
<https://doi.org/10.1080/09603123.2021.1905782>.

Montesinos López, O.A., Montesinos López, A. and Crossa, J. (2022) *Multivariate Statistical Machine Learning Methods for Genomic Prediction, Multivariate Statistical Machine Learning Methods for Genomic Prediction*. Springer International Publishing. Available at:
<https://doi.org/10.1007/978-3-030-89010-0>.

Moonasar, D.M. *et al.* (2021) 'Progress, challenges and priorities for malaria elimination in South Africa', *Transactions of the Royal Society of South Africa*, 76(2), pp. 105–116. Available at: <https://doi.org/10.1080/0035919X.2021.1890274>.

Mughal, H. *et al.* (2022) 'Random Forest Model Predictions Afford Dual-Stage Antimalarial Agents', *ACS Infectious Diseases*, 8(8), pp. 1553–1562. Available at:
<https://doi.org/10.1021/acsinfecdis.2c00189>.

- Musetha, M.A. (2016) 'The impact of climate change on agricultural crop production in the Vhembe District Municipality, Limpopo Province South Africa', *Plant Genomics and Climate Change*, pp. 1–13. Available at: https://doi.org/10.1007/978-1-4939-3536-9_1.
- Mussumeci, E. and Codeço Coelho, F. (2020) 'Large-scale multivariate forecasting models for Dengue - LSTM versus random forest regression', *Spatial and Spatio-temporal Epidemiology*, 35, p. 100372. Available at: <https://doi.org/10.1016/j.sste.2020.100372>.
- Mwagha, S.M. (2016) *A VISUAL WEATHER LORE VERIFICATION TOOL USING FUZZY COGNITIVE MAPS BASED ON COMPUTER VISION*.
- Mwagha, S.M. and Masinde, M. (2015) *Scientific Verification of Weather Lore for Drought Forecasting-The Role of Fuzzy Cognitive Mapping*. Available at: www.IST-Africa.org/Conference2015.
- Mzelikahle, K. and Mapuma, D.J. (2018) 'Optimisation of Self Organising Maps Using the Bat Algorithm Optimisation of Self Organising Maps Using the Bat Algorithm', (December).
- Nápoles, G. *et al.* (2018) 'FCM Expert: Software Tool for Scenario Analysis and Pattern Classification Based on Fuzzy Cognitive Maps', *International Journal on Artificial Intelligence Tools*, 27(7). Available at: <https://doi.org/10.1142/S0218213018600102>.
- Naqvi, S.R. *et al.* (2018) 'Artificial neural networks based dynamic priority arbitration for asynchronous flow control', *Neural Computing and Applications*, 29(7), pp. 627–637. Available at: <https://doi.org/10.1007/s00521-016-2571-6>.
- National Department of Health South Africa (2018) 'MALARIA-ELIMINATION-STRATEGIC-PLAN-FOR-SOUTH-AFRICA-2019-2023-MALARIA-ELIMINATION-STRATEGIC-PLAN-2019-2023 (1)'.
- National Institute for Communicable Disease (2019) 'NCID Disease Index- Malaria'. Available at: <https://www.nicd.ac.za/diseases-a-z-index/malaria/> (Accessed: 23 September 2023).
- National Institute for Communicable Diseases (2023) *NICD DISEASE INDEX – MALARIA*.
- Nenwiini, S., Kabanda, T. and Bokone-bophirima, Y.Y. (2017) *Climatic Anomalies and their influence on Rainfall Trends in Vhembe District South Africa*.
- Nkiruka, O., Prasad, R. and Clement, O. (2021) 'Prediction of malaria incidence using climate variability and machine learning', *Informatics in Medicine Unlocked*, 22, p. 100508. Available at: <https://doi.org/10.1016/j.imu.2020.100508>.
- Nozari, M.A. *et al.* (2021) 'Analysis of the sharing economy effect on sustainability in the transportation sector using fuzzy cognitive mapping', *Journal of Cleaner Production*, 311. Available at: <https://doi.org/10.1016/j.jclepro.2021.127331>.

Nunamaker, J.F., Chen, M. and Purdin, T.D.M. (1990) 'Systems Development in Information Systems Research', *Journal of Management Information Systems*, 7(3), pp. 89–106. Available at: <https://doi.org/10.1080/07421222.1990.11517898>.

Nyetanyane, J. (2023) 'Indigenous Knowledge Mobile Based Application that Quantifies Farmers' Season Predictions with the Help of Scientific Knowledge', in M. Masinde and A. Bagula (eds) *Emerging Technologies for Developing Countries. AFRICATEK 2022. Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering*, pp. 192–205. Available at: https://doi.org/10.1007/978-3-031-35883-8_13.

Nyetanyane, J. and Masinde, M. (2020) 'Integration of Indigenous Knowledge, Climate Data, Satellite Imagery and Machine Learning to Optimize Cropping Decisions by Small-Scale Farmers. a Case Study of uMgungundlovu District Municipality, South Africa', in *Lecture Notes of the Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering, LNICST*. Springer, pp. 3–19. Available at: https://doi.org/10.1007/978-3-030-51051-0_1.

Nyong, A., Adesina, F. and Osman Elasha, B. (2007) 'The value of indigenous knowledge in climate change mitigation and adaptation strategies in the African Sahel', *Mitigation and Adaptation Strategies for Global Change*, 12(5), pp. 787–797. Available at: <https://doi.org/10.1007/s11027-007-9099-0>.

Ong, S.Q. *et al.* (2023) 'Predicting dengue transmission rates by comparing different machine learning models with vector indices and meteorological data', *Scientific Reports*, 13(1). Available at: <https://doi.org/10.1038/s41598-023-46342-2>.

Onwuegbuzie, R.B.J. and A.J. (2014) 'Mixed Methods Research: A Research Paradigm Whose Time Has Come.', *American Educational Research Association*, 33(7), pp. 14–26.

Onyijen, O.H., Ogieriakhi, O. and Olaitan, E.O. (2023) *MALARIA DISEASE PREDICTION IN WEST AFRICA USING SELECTED MACHINE LEARNING TECHNIQUE*, *Western European Journal of Medicine and Medical Science*. Available at: <https://westerneuropeanstudies.com/index.php/3https://westerneuropeanstudies.com/index.php/3>.

Orang, O., de Lima e Silva, P.C. and Guimarães, F.G. (2022) 'Time series forecasting using fuzzy cognitive maps: a survey', *Artificial Intelligence Review* [Preprint]. Available at: <https://doi.org/10.1007/s10462-022-10319-w>.

Orlowska, E. (2013) *Incomplete information: Rough set analysis*. Physica-Verlag.

Östlund, U. *et al.* (2011) 'Combining qualitative and quantitative research within mixed method research designs: A methodological review', *International Journal of Nursing Studies*, pp. 369–383. Available at: <https://doi.org/10.1016/j.ijnurstu.2010.10.005>.

Panch, T., Szolovits, P. and Atun, R. (2018) 'Artificial intelligence, machine learning and health systems', *Journal of Global Health*, 8(2), pp. 1–8. Available at: <https://doi.org/10.7189/jogh.08.020303>.

Papageorgiou, E. *et al.* (2018) *A web-based tool for Fuzzy Cognitive Map Modeling*. Available at: <https://scholarsarchive.byu.edu/iemssconferencehttps://scholarsarchive.byu.edu/iemssconference/2018/Stream-C/73>.

Papageorgiou, E.I. *et al.* (2009) *2009 IEEE International Conference on Fuzzy Systems : proceedings : ICC Jeju, Jeju Island, Korea : August 20-24, 2009*. [IEEE].

Parbat, D. and Chakraborty, M. (2020) 'A python based support vector regression model for prediction of COVID19 cases in India', *Chaos, Solitons & Fractals*, 138, p. 109942. Available at: <https://doi.org/10.1016/j.chaos.2020.109942>.

Patrick, S.M. *et al.* (2023) 'Household living conditions and individual behaviours associated with malaria risk: a community-based survey in the Limpopo River Valley, 2020, South Africa', *Malaria Journal*, 22(1). Available at: <https://doi.org/10.1186/s12936-023-04585-4>.

Pauna, V.H. *et al.* (2018) 'The scientific research on ecosystem services: a bibliometric analysis.', *Ecological Questions*, 29(3), p. pp.53-62.

Peiffer-Smadja, N. *et al.* (2020) 'Machine learning for clinical decision support in infectious diseases: a narrative review of current applications', *Clinical Microbiology and Infection*, 26(5), pp. 584–595. Available at: <https://doi.org/10.1016/j.cmi.2019.09.009>.

Phoobane, P. and Masinde, M. (2023) 'Investigating the adoption of indigenous knowledge in mitigating climate-linked challenges', *International Journal of Research in Business and Social Science* (2147- 4478), 12(7), pp. 394–404. Available at: <https://doi.org/10.20525/ijrbs.v12i7.2728>.

Phoobane, P., Masinde, M. and Botai, J. (2022) *Prediction Model for Malaria: An Ensemble of Machine Learning and Hydrological Drought Indices, Lecture Notes in Networks and Systems*. Available at: https://doi.org/10.1007/978-981-16-1781-2_51.

Phoobane, P., Masinde, M. and Mabhaudhi, T. (2022) 'Predicting Infectious Diseases: A Bibliometric Review on Africa', *International Journal of Environmental Research and Public Health*. MDPI. Available at: <https://doi.org/10.3390/ijerph19031893>.

Pisner, D.A. and Schnyer, D.M. (2020) 'Support vector machine', in *Machine Learning*. Elsevier, pp. 101–121. Available at: <https://doi.org/10.1016/B978-0-12-815739-8.00006-7>.

Ponce, H. and Gutiérrez, S. (2019) 'An indoor predicting climate conditions approach using Internet-of-Things and artificial hydrocarbon networks', *Measurement: Journal of the International Measurement Confederation*, 135, pp. 170–179. Available at: <https://doi.org/10.1016/j.measurement.2018.11.043>.

- Popp, J. and Ortloff, D. (2015) 'High-Tech R & D – Drowning in', (July 2012), pp. 9–12.
- Pradhan, N. *et al.* (2020) 'Diabetes prediction using artificial neural network', *Deep Learning Techniques for Biomedical and Health Informatics*, 121, pp. 327–339. Available at: <https://doi.org/10.1016/b978-0-12-819061-6.00014-8>.
- Preece, J., Rogers, Y. and Sharp, H. (2015) *IINTERACTION DESIGN beyond humancomputer interaction*. 4th ed. United Kingdom: John Wiley & Sons Ltd.
- Preethi, P.R.A. and Elavarasi, P.J. (2017) 'Big Data Analytics Using Hadoop Tools – Apache Hive Vs Apache Pig', 24(3), pp. 16–20.
- Quansah, J.E., Engel, B. and Rochon, G.L. (2010) 'Early Warning Systems: A Review', *Journal of Terrestrial Observation*, 2(2).
- Rahayu, H.P. *et al.* (2020) 'A study of people-centered early warning system in the face of near-field tsunami risk for Indonesian coastal cities', *International Journal of Disaster Resilience in the Built Environment*, 11(2), pp. 241–262. Available at: <https://doi.org/10.1108/IJDRBE-10-2019-0068>.
- Ritchie, A., Teufel, S. and Robertson, S. (2008) 'Using terms from citations for IR: Some first results', *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 4956 LNCS(April), pp. 211–221. Available at: https://doi.org/10.1007/978-3-540-78646-7_21.
- Roopa, S. and Rani, M. (2012) 'Questionnaire Designing for a Survey', *The Journal of Indian Orthodontic Society*, 46, pp. 273–277. Available at: <https://doi.org/10.5005/jp-journals-10021-1104>.
- Sabra, S. *et al.* (2020) 'A hybrid knowledge and ensemble classification approach for prediction of venous thromboembolism', *Expert Systems*, 37(1), pp. 1–14. Available at: <https://doi.org/10.1111/exsy.12388>.
- Sadek, R.M. *et al.* (2019) 'Parkinson's Disease Prediction Using Artificial Neural Network', 3(1), pp. 1–8.
- Sadoine, M.L. *et al.* (2024) 'Predicting malaria risk considering vector control interventions under climate change scenarios', *Scientific Reports*, 14(1). Available at: <https://doi.org/10.1038/s41598-024-52724-x>.
- Sailasya, G. and Kumari, G.L.A. (2021) 'Analyzing the Performance of Stroke Prediction using ML Classification Algorithms', *International Journal of Advanced Computer Science and Applications*, 12(6). Available at: <https://doi.org/10.14569/IJACSA.2021.0120662>.
- Salam, H.A. and Khan, B.M. (2016) 'Use of wireless system in healthcare for developing countries', *Digital Communications and Networks*, 2(1), pp. 35–46. Available at: <https://doi.org/10.1016/j.dcan.2015.11.001>.

- Sarkar, B.K. (2019) 'An e-healthcare system for disease prediction using hybrid data mining technique', pp. 628–661. Available at: <https://doi.org/10.1108/JM2-05-2018-0069>.
- Saunders, M., Lewis, P. and Thornhill, A. (2009) *Research methods for business students*. Pearson education.
- Scheinman, D. (2000) 'An integrated program for developing medicinal plants: A case study from Tanga, Tanzania.', in *In Medicinal Plants Forum for Commonwealth Africa*. Cape Town, pp. 4–6.
- Seeram, E. (2019) 'An overview of correlational research', *Radiologic technology*, 91(2), pp. 176–179.
- Segun, O.E. *et al.* (2020) 'Statistical Modelling of the Effects of Weather Factors on Malaria Occurrence in Abuja, Nigeria', *International Journal of Environmental Research and Public Health*, 17(10), p. 3474. Available at: <https://doi.org/10.3390/ijerph17103474>.
- Sharma, G. (2017) 'Impact Factor: 5.2 Pros and cons of different sampling techniques', *International Journal of Applied Research*, 3(7), pp. 749–752. Available at: www.allresearchjournal.com.
- Sharma, V. *et al.* (2015) 'Malaria Outbreak Prediction Model Using Machine Learning', *International Journal of Advanced Research in Computer Engineering & Technology*, 4(12), pp. 4415–4419.
- Sheppard, S.R.J. and Cizek, P. (2009) 'The ethics of Google Earth: Crossing thresholds from spatial data to landscape visualisation', *Journal of Environmental Management*, 90(6), pp. 2102–2117. Available at: <https://doi.org/10.1016/j.jenvman.2007.09.012>.
- Sheykhmousa, M. *et al.* (2020) 'Support Vector Machine Versus Random Forest for Remote Sensing Image Classification: A Meta-Analysis and Systematic Review', *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*. Institute of Electrical and Electronics Engineers Inc., pp. 6308–6325. Available at: <https://doi.org/10.1109/JSTARS.2020.3026724>.
- Shi, Y. *et al.* (2016) 'Three-Month Real-Time Dengue Forecast Models: An Early Warning System for Outbreak Alerts and Policy Decision Support in Singapore', *Environmental Health Perspectives*, 124(9), pp. 1369–1375. Available at: <https://doi.org/10.1289/ehp.1509981>.
- Singh, V. *et al.* (2020) 'Prediction of COVID-19 corona virus pandemic based on time series data using support vector machine', *Journal of Discrete Mathematical Sciences and Cryptography*, 23(8), pp. 1583–1597. Available at: <https://doi.org/10.1080/09720529.2020.1784535>.
- Smylie, J. *et al.* (2004) 'Knowledge translation and indigenous knowledge.', *International journal of circumpolar health*, 63 Suppl 2, pp. 139–143. Available at: <https://doi.org/10.3402/ijch.v63i0.17877>.

- Smylie, J., Kaplan-Myrth, N. and Mcshane, K. (2009) 'Indigenous Knowledge Translation: Baseline Findings in a Qualitative Study of the Pathways of Health Knowledge in Three Indigenous Communities in Canada', *Health Promotion Practice*, 10(3), pp. 436–446. Available at: <https://doi.org/10.1177/1524839907307993>.
- Somani, A.K. and Deka, G.C. eds (2017) *Big data analytics: Tools and technology for effective planning*. CRC Press.
- Son, H.N., Chi, D.T.L. and Kingsbury, A. (2019) 'Indigenous knowledge and climate change adaptation of ethnic minorities in the mountainous regions of Vietnam: A case study of the Yao people in Bac Kan Province', *Agricultural Systems*, 176(July). Available at: <https://doi.org/10.1016/j.agsy.2019.102683>.
- Soropa, G. *et al.* (2015) 'Indigenous knowledge system weather forecasts as a climate change adaptation strategy in smallholder farming systems of Zimbabwe: Case study of Murehwa, Tsholotsho and Chiredzi districts', *African Journal of Agricultural Research*, 10(10), pp. 1067–1075. Available at: <https://doi.org/10.5897/ajar2013.7205>.
- South Africa Weather Service (2022) *Organization Overview*. Available at: <https://www.weathersa.co.za/> (Accessed: 25 September 2023).
- Statista (2024) *Worldwide digital population*.
- Steimle, L.N. and Denton, B.T. (2017) 'Screening and Treatment of Chronic Diseases', *Markov Decision Processes in Practice*, pp. 1–33. Available at: https://doi.org/10.1007/978-3-319-47766-4_6.
- Stratton, S.J. (2021) 'Population Research: Convenience Sampling Strategies', *Prehospital and Disaster Medicine*. Cambridge University Press, pp. 373–374. Available at: <https://doi.org/10.1017/S1049023X21000649>.
- Studer, S. *et al.* (2021) 'Towards CRISP-ML(Q): A Machine Learning Process Model with Quality Assurance Methodology', *Machine Learning and Knowledge Extraction*, 3(2), pp. 392–413. Available at: <https://doi.org/10.3390/make3020020>.
- Sullivan, J.H., Warkentin, M. and Wallace, L. (2021) 'So many ways for assessing outliers: What really works and does it matter?', *Journal of Business Research*, 132, pp. 530–543. Available at: <https://doi.org/10.1016/j.jbusres.2021.03.066>.
- Symposium, D., Engineering, I. and Morais, S. (2015) 'Survey on Frameworks for Distributed Computing: Hadoop, Spark and Storm', *Proceedings of the 10th Doctoral Symposium in Informatics Engineering - DSIE'15*, pp. 95–105.
- Tanner, L. *et al.* (2008) 'Decision tree algorithms predict the diagnosis and outcome of dengue fever in the early phase of illness', *PLoS Neglected Tropical Diseases*, 2(3). Available at: <https://doi.org/10.1371/journal.pntd.0000196>.

- Tashakkori, A. and Teddlie, C. (2009) *Foundations of mixed methods research: Integrating quantitative and qualitative approaches in the social and behavioral sciences*. SAGE publications.
- Tayefi, M. *et al.* (2017) 'hs-CRP is strongly associated with coronary heart disease (CHD): A data mining approach using decision tree algorithm', *Computer Methods and Programs in Biomedicine*, 141, pp. 105–109. Available at: <https://doi.org/10.1016/j.cmpb.2017.02.001>.
- Thakur, S. and Dharavath, R. (2019) 'Artificial neural network based prediction of malaria abundances using big data: A knowledge capturing approach', *Clinical Epidemiology and Global Health*, 7(1), pp. 121–126. Available at: <https://doi.org/10.1016/j.cegh.2018.03.001>.
- Theocharides, S. *et al.* (2018) 'Machine learning algorithms for photovoltaic system power output prediction', *2018 IEEE International Energy Conference, ENERGYCON 2018*, (October), pp. 1–6. Available at: <https://doi.org/10.1109/ENERGYCON.2018.8398737>.
- Thomas, G. (2017) *How to do your research project: A guide for students*. 3rd editio. London: Sage.
- Thomson, M.C. *et al.* (2018) 'Climate drivers of vector-borne diseases in Africa and their relevance to control programmes', *Infectious Diseases of Poverty*, 7(1), pp. 1–22. Available at: <https://doi.org/10.1186/s40249-018-0460-1>.
- Tiu, L.A. *et al.* (2021) 'Literature Review: Impact Of Temperature And Rainfall On Incident Malaria', *IOP Conference Series: Earth and Environmental Science*, 755(1), p. 012084. Available at: <https://doi.org/10.1088/1755-1315/755/1/012084>.
- Tongco, M.D.C. (2007) 'Purposive sampling as a tool for informant selection. Ethnobotany research and applications', *Ethnobotany Research and Applications*, 5, pp. 147–158. Available at: <https://doi.org/10.17348/era.5.0.147-158>.
- Tsatsaris, A. *et al.* (2021) 'Geoinformation Technologies in Support of Environmental Hazards Monitoring under Climate Change: An Extensive Review', *ISPRS International Journal of Geo-Information*, 10(2), p. 94. Available at: <https://doi.org/10.3390/ijgi10020094>.
- Tsuji, L.J. and Ho, E. (2002) 'Traditional environmental knowledge and western science: in search of common ground', *Canadian Journal of Native Studies*, 22(2), pp. 327–360.
- UNISDR (2006) 'Developing Early Warning Systems : A Checklist', *Third International Conference on Early Warning*, (March), pp. 1–13.
- Vera-Polania, F. *et al.* (2014) 'Bibliometric assessment of the Latin-American contributions in dengue', *Recent Patents on Anti-Infective Drug Discovery*, 9(3), pp. 195–201.
- Verma, A.K. *et al.* (2020) 'A new backpropagation neural network classification model for prediction of incidence of malaria', *Frontiers in Bioscience*, 25(2), p. 4808. Available at: <https://doi.org/10.2741/4808>.

- Vhembe District Municipality (2024) *Vhembe District Municipality (DC34)*. Available at: <https://municipalities.co.za/map/129/vhembe-district-municipality> (Accessed: 26 September 2024).
- Vijander, S. *et al.* (2020) 'Prediction of COVID-19 corona virus pandemic based on time series data using Support Vector Machine', *Journal of Discrete Mathematical Sciences and Cryptography*, 23(8), pp. 1583–1597.
- Vijayashree, J. and Iyengar, N.C.S.N. (2016) 'Heart disease prediction system using data mining and hybrid intelligent techniques: A review', *International Journal of Bio-Science and Bio-Technology*, 8(4), pp. 139–148. Available at: <https://doi.org/10.14257/ijbsbt.2016.8.4.16>.
- Vujovic, Ž.Đ. (2021) 'Classification Model Evaluation Metrics', *International Journal of Advanced Computer Science and Applications*, 12(6). Available at: <https://doi.org/10.14569/IJACSA.2021.0120670>.
- Wallisch, C. *et al.* (2022) 'Review of guidance papers on regression modeling in statistical series of medical journals', *PLOS ONE*, 17(1), p. e0262918. Available at: <https://doi.org/10.1371/journal.pone.0262918>.
- Wang, M. *et al.* (2022) 'Deep learning model for multi-classification of infectious diseases from unstructured electronic medical records', *BMC Medical Informatics and Decision Making*, 22(1), p. 41. Available at: <https://doi.org/10.1186/s12911-022-01776-y>.
- Wang, S., Cao, J. and Yu, P.S. (2022) 'Deep Learning for Spatio-Temporal Data Mining: A Survey', *IEEE Transactions on Knowledge and Data Engineering*, 34(8), pp. 3681–3700. Available at: <https://doi.org/10.1109/TKDE.2020.3025580>.
- Wang, Y., Kung, L.A. and Byrd, T.A. (2018) 'Big data analytics: Understanding its capabilities and potential benefits for healthcare organizations', *Technological Forecasting and Social Change*, 126, pp. 3–13. Available at: <https://doi.org/10.1016/j.techfore.2015.12.019>.
- Wilhite, D.A. and Glantz, M.H. (1985) 'Understanding: the Drought Phenomenon: The Role of Definitions', *Water International*, 10(3), pp. 111–120. Available at: <https://doi.org/10.1080/02508068508686328>.
- World Health Organisation (2017) 'WHO | Plague–Madagascar'. Available at: <https://journals.plos.org/plosntds/article?id=10.1371/journal.pntd.0006131>.
- World Health Organisation (2019) 'Countries slides Measles'.
- World Health Organisation (2021) *WHO Coronavirus Disease (COVID-19) Dashboard*. Available at: <https://covid19.who.int/>.
- World Health Organization (2018) 'Global Health estimate 2016: Deaths by Cause, Age, Sex by Country and Region, 2000-2016.', *Geveva, World Health Organisation* [Preprint].

World Health Organization (2019) *World Malaria Report 2019*, Geneva: World Health Organization. Geneva:

World Health Organization (2020a) 'Top 10 causes of death'. Available at: <https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death> (Accessed: 30 August 2024).

World Health Organization (2020b) *Top 10 Causes of Death*.

World Health Organization (2022) *World malaria report 2022*. Available at: <https://www.who.int/teams/global-malaria-programme/reports/world-malaria-report-2022> (Accessed: 7 September 2024).

World Health Organization (2023) *Malaria*, <https://www.who.int/news-room/fact-sheets/detail/malaria>.

World Health Organization African Regional Office (2014) 'The African Regional Health Report: The Health of the People', *World Health Organization*, p. 196.

Woyessa, A. *et al.* (2023) 'El Niño and other climatic drivers of epidemic malaria in Ethiopia: new tools for national health adaptation plans', *Malaria Journal*, 22(1), p. 195. Available at: <https://doi.org/10.1186/s12936-023-04621-3>.

Yadav, M., Perumal, M. and Srinivas, M. (2020) 'Analysis on novel coronavirus (COVID-19) using machine learning methods', *Chaos, Solitons and Fractals*, 139. Available at: <https://doi.org/10.1016/j.chaos.2020.110050>.

Yang, D. *et al.* (2021) 'Detection and analysis of COVID-19 in medical images using deep learning techniques', *Scientific Reports*, 11(1). Available at: <https://doi.org/10.1038/s41598-021-99015-3>.

Yang, W., Zhang, J. and Ma, R. (2020) 'The prediction of infectious diseases: A bibliometric analysis', *International Journal of Environmental Research and Public Health*, 17(17), pp. 1–19. Available at: <https://doi.org/10.3390/ijerph17176218>.

Yeşilkanat, C.M. (2020) 'Spatio-temporal estimation of the daily cases of COVID-19 in worldwide using random forest machine learning algorithm', *Chaos, Solitons & Fractals*, 140, p. 110210. Available at: <https://doi.org/10.1016/j.chaos.2020.110210>.

Yilmaz, K. (2013) *Comparison of Quantitative and Qualitative Research Traditions: epistemological, theoretical, and methodological differences*, *European Journal of Education*.

Yinka-Ogunleye, A. *et al.* (2018) 'Reemergence of human monkeypox in Nigeria, 2017.', *Emerging infectious diseases*, 24(6), p. p.1149.

York University Libraries (2024) *Limitations of bibliometrics*.

Zahra, A. *et al.* (2023) 'Simulation of the potential impact of climate change on malaria incidence using artificial neural networks (ANNs)', *Environmental Science and Pollution Research*, pp. 1–20.

Zhang, S. *et al.* (2019) 'Deep learning based recommender system: A survey and new perspectives', *ACM Computing Surveys*, 52(1). Available at: <https://doi.org/10.1145/3285029>.

Zhao, N. *et al.* (2020) 'Machine learning and dengue forecasting: Comparing random forests and artificial neural networks for predicting dengue burden at national and sub-national scales in Colombia', *PLoS Neglected Tropical Diseases*, 14(9), pp. 1–16. Available at: <https://doi.org/10.1371/journal.pntd.0008056>.

Zheng, Z. *et al.* (2015) 'Real-time big data processing framework: challenges and solutions', *Appl. Math. Inf. Sci.*, 9(6), p. 3169.

Appendix A: IK Indicators for Malaria Forecast in SSA

Insects IK indicators for malaria prediction based on rainfall and temperature prediction.

Insects		
Indicators	Significance	Reference
Termite (<i>Ancistrotermes</i> spp)	The appearance of many termites indicates the onset of rainfall. When they carry grass and rapid movements that signifies imminent rainfall. However, these observations do not indicate the intensity of the rain and do not show the level of malaria season.	(Kijazi, Chang'a, Liwenga, 2013; Macherera, Chimbari and Mukaratirwa, 2017; Mafongoya, and Ajayi, 2017)
Beetle (cleoptera)	When the beetle hits the ground with its tail as a sign of good rain in the year. This usually happens around September and October. This shows that there is likely to be a high incidence of malaria.	(Macherera et al. 2017)

Butterflies	White butterflies generally fly from west to east in large numbers signifying that much rain is expected. If the butterflies are observed in October. It indicates that the season is likely to be a good one in terms of rain abundance but implies that there would be high malaria incidence. Butterfly movement from west to east is an indication of imminent rainfall. The occurrence and migration of butterflies from the south to the north is an indication of the early onset of rains and also gives a prospect of a good season. Specifically, black butterflies indicate a prospect of a good season.	(Macherera, Chimbari and Mukaratirwa, 2017; Mafongoya, and Ajayi, 2017)
Ants	A lot of red ants rapidly increasing the size of anthills means good rainfall can be expected. The appearance of army ants in large numbers almost everywhere even inside the houses and appear to be celebrating is an indication of the onset of rains and mostly heavy rains (only about a week before onset). Occurrence of flying ants after strong sunshine indicates proper rains have started.	(Mafongoya, Jiri, Mubaya, and Mafongoya, 2017) (Mahoo et al., 2015)
Bees	Bees appearing in big groups indicate long rainy season is about to start.	(Mahoo et al. 2015)
Bee eater (Merops hirundineus)	Appearance in October indicates imminent rainfall and/or onset.	(Mafongoya, and Ajayi, 2017)
Cicadas, day-flying chafers, dragon flies	The appearance of these insects indicates imminent rainfall.	
Millipedes, spiders	The appearance of millipedes, spiders and certain insects indicates the coming of heavy rainfalls.	(Kolawole et al., 2016; Mafongoya, and Ajayi, 2017)
Thrips	When these insects are seen surrounding wetland areas, it is an indication of a good rainy season.	(Mahoo et al. 2015)
Worms	The appearance of many armyworms (Spodoptera exempta) on trees in October means good rains in the coming season. Abundant appearance of worms just after the first rain indicates a wet year.	(Kijazi, Chang'a, Liwenga, 2013; Mahoo et al., 2015; Soropa et al., 2015; Mafongoya, and Ajayi, 2017)
Grasshopper (Hesperotettix speciosus)	The appearance of many grasshoppers means less rainfall. The appearance of green grasshoppers in large numbers in the fields indicates the onset of rainfall especially short season rains.	(Kijazi, Chang'a, Liwenga, 2013; Mahoo et al., 2015; Mafongoya, and Ajayi, 2017)

Scorpion (Arachnida spp)	The appearance of black or dark scorpions in October indicates the possibility of a good rainy season.	(Kijazi, Chang'a, Liwenga, 2013; Mafongoya, and Ajayi, 2017)
-----------------------------	--	--

Birds IK indicators for malaria prediction based on rainfall and temperature prediction.

Birds		
Indicators	Significance	Reference
Guinea fowl	Laying of many eggs by guinea fowls is a sign of a good rainy season and that also implies high malaria incidence would be observed.	(Macherera et al. 2017)
Coucal (Centropus sp) and wild birds	Singing out loud is an indication of short rains.	(Mahoo et al. 2015)
Ostrich	When ostriches run with their wings flapping their bodies make noise that sounds like that of a drum that signifies a good rainy season ahead. This also means a high incidence of malaria. The presence of ostriches roaming in the vicinity of settlements indicates no rainfall.	(Kolawole et al. 2016; Macherera et al. 2017)
Swallow and sparrows	Large numbers of swallows observed flying in the sky signify of coming of rain and if that happens in November that means the onset of rains will be imminent. However, this observation does not provide information on the nature of the malaria season to be expected, rather it only shows that the malaria season is imminent. First appearance of sparrows; a flock of swallows preceding dark clouds means imminent rainfall, sometimes within 2-3 days.	(Kijazi, Chang'a, Liwenga, 2013; Mahoo et al., 2015; Macherera, Chimbari and Mukaratirwa, 2017; Mafongoya, and Ajayi, 2017)
Ground hornbill	Ground hornbill makes noise around September to November if there is going to be a good rainy season and that indicates the beginning of malaria season.	(Macherera et al. 2017)
Cattle egret birds (Bubulcus ibis)	The appearance of cattle egret birds in October indicates imminent rainfall and a good rainy season.	(Kijazi, Chang'a, Liwenga, 2013; Mafongoya, and Ajayi, 2017)
Migratory birds Hornbills	The presence of large numbers of migratory birds, hornbills and certain bird species signifies the approaching of the rainy season and that indicates the beginning of malaria season.	(Mahoo et al. 2015; Kolawole et al. 2016; Macherera et al. 2017)
Stock	The appearance of stock and certain birds singing, nesting, and chirping of certain birds indicate rainfall is near.	(Soropa <i>et al.</i> , 2015; Mafongoya and Ajayi, 2017)

Coucal birds (Centropus spp) Golden oriole	The singing of these birds early in the morning indicates the onset of rain and a good rainy season.	(Kijazi, Chang'a, Liwenga, 2013; Mahoo et al., 2015; Mafongoya, and Ajayi, 2017)
Pangolin (scaly ant-eaters)	Appearance indicates imminent rainfall occurrence.	(Kijazi, Chang'a , Liwenga , 2013; Mafongoya, and Ajayi, 2017)
Ducks	Ducks stretching their wings and playing in the dust is a sign of the onset of rain, especially during the short rainy season	(Mahoo et al. 2015)
Chicken	Chicken stretching their wings repeatedly is a sign of the onset of short rains.	(Mahoo et al. 2015)
Fischer's Lovebirds	When these birds are heard crying is an indication of short rains.	(Mahoo et al. 2015)
Wild birds	The appearance of certain wild birds walking to the valley or wet lands is an indication of short rains while the presence of certain wild birds indicates long rains.	(Mahoo et al. 2015)

Animal IK indicators for malaria prediction

Animals		
Indicators	Significance	Reference
Wild animals	Wild animals bearing much offspring is an indicator of a good rainy season and high malaria incidence.	(Macherera et al. 2017)
Lions Baboons, monkeys, leopards and antelopes	Lions passing by the home settlements without attacking livestock or people around September and October are a sign of a good rainy season ahead and a high incidence of malaria. Baboons, monkeys, leopards and antelopes passing by the village is a sign of a good rainy season ahead. Heavy sounds of lions and other wild animals around October are an indication of the rainy season.	(Mahoo et al. 2015; Soropa et al. 2015; Macherera et al. 2017)
Sheep, goats, cattle	Ear flapping by these animals is an indication of the onset and prospects of a good season. Calves jumping happily means good rain seasons.	(Mahoo et al., 2015; Mafongoya, and Ajayi, 2017)
Pigs	Grunting pigs indicate imminent rainfall.	(Mafongoya, and Ajayi, 2017)
Frogs	When frogs croak in swampy areas indicates the beginning of the rainy season. Frogs/African clawed frogs starting to make a lot of noise is an indication of rainfall onset and the louder the noises become it signifies a good season	(Mahoo et al., 2015; Mafongoya, Pand Ajayi, 2017)

Snails	Appearance of shelled and garden snails in large numbers especially at the time they are not expected is an indication of the onset of a short rain season and a prospect of a good season	
Snakes	Certain snakes moving down the mountain are an indication of a good rain season.	(Mafongoya and Ajayi, 2017)
Rabbit	When a rock rabbit makes an unusual squeak, it indicates imminent rainfall.	(Mafongoya and Ajayi, 2017)
Tortoises	Their frequent appearance indicates good rain seasons.	(Mafongoya, and Ajayi, 2017)
Springhare Ant-bear	When springhare and ant-bear are spotted during the day, this is an indication of no rain.	(Kolawole et al. 2016)

Plants IK indicators for malaria prediction based on rainfall and temperature prediction.

Plants		
Indicators	Significance	Reference
Lanchocarpus capassa Rolfe	If this tree produces flowers around October and November, that may mean that there would be much rain in the following rainy season and that malaria incidence would be expected to be high. If this tree flowers early and produces many fruits by November, much rain will fall during that season. This also indicates that the incidence of malaria will be high	(Macherera et al. 2017)
Albizia trees	Their presence on Albizia trees with water dripping from them is an indication of a good season (almost one month before the rains start).	(Mahoo et al. 2015)
Boscia Albitrunca (Burch.) Gilg Benedict Acacia karroo Hayne Shepherd's tree	When this tree bears many fruits around October and November, there will be low rainfall in that season and the incidence of malaria will be expected to be low in that season. However, when a shepherd's tree bears an abundance of fruit is an indication of heavy rainfall.	(Kolawole et al. 2016; Macherera et al. 2017)
Colophospermum mopane (Benth) Leonard	When this tree flowers and bears fruit before the rain falls, there will be low rainfall during that season and the malaria incidence will be expected to be low	(Macherera et al. 2017)
Cassia Abbreviata Oliv	When Cassia abbreviata produces new leaves during October and November that rainy season will be expected to be good and much malaria would be expected.	(Macherera et al. 2017)

Sclerocary abirrea (A Rich.) Hochst	The flowering of this tree around October signifies a good rainy season and a high incidence of malaria. If it flowers later than October less rain will be expected and the same for malaria incidence.	(Macherera et al. 2017)
Jasminum stenolobum Rolfe	Early flowering of this tree signifies good rains for the year and a high incidence of malaria.	(Macherera et al. 2017)
Afzelia quanzensis Welw	If this tree produces new leaves early around September and October, indicates a good rainy season that year and high malaria incidence	(Macherera et al. 2017)
Acacia nigrescens Oliv	If the tree bears a lot of fruits it means that little rain is expected. If there is drought the old leaves do not fall so the tree delays producing new Leaves. Malaria incidence will be expected to be low.	(Macherera et al. 2017)
Peach tree, apricot and acacia species	The flowering of the peach tree (<i>Prunus persica</i>), apricot (<i>Prunus armeniaca</i>) and the budding of acacia species indicate the beginning of the rainy season.	(Mafongoya and Ajayi, 2017)
Coffee, delonix and mango trees; and other trees	The flowering of coffee, delonix, mango trees and other trees does not indicate the rainy season, rather it means it will rain soon. However, if mango trees' leaves sprout, it indicates the approach of the rain season.	(Mafongoya and Ajayi, 2017)
Behaviour of certain plants: sprouting of Aloe ferox;	Sprouting of Aloe ferox; germination of new leaves on baobab and tamarind trees indicates good rains Drying of flowers and early wilting and exfoliation of (Kolawole <i>et al.</i> , 2016; Mafongoya and Ajayi, 2017)2016; Mafongoya and Ajayi, 2017)	
Muchakata, gan'acha, mushuku and mango trees Pears, coffee, avocado and plums trees, Loquarts and pigeon pea plant	Heavy flowering of Muchakata, gan'acha, mushuku and mango trees indicates a potential drought season while flowering of Pears, coffee, avocado and plums trees, Loquarts and pigeon pea plants indicates rain onset and a good rainy season. Early flowering of certain trees indicates more rainfall while late flowering indicates drought season. Dormancy breaking in certain tree species e.g. mupfuti (<i>Brachystegia boehmii</i>) indicates plenty of rain in a few days Profuse fruiting of certain tree species indicates a challenging rainy season. Failure of these trees to flower indicates no rain.	(Mafongoya <i>et al.</i> , 2017) (Kolawole <i>et al.</i> , 2016) (Kijazi <i>et al.</i> , 2013) (Kijazi, Chang'a and Liwenga, 2013; Mahoo <i>et al.</i> , 2015; Soropa <i>et al.</i> , 2015; Kolawole <i>et al.</i> , 2016; Mafongoya and Ajayi, 2017)
Datura	Abundant fruit production and dropping leaves of the Datura an indications of rainfall onset.	(Mahoo <i>et al.</i> 2015)
Wild fruits	Plenty of wild fruits in October and November indicate the rainy season.	(Soropa <i>et al.</i> 2015)

Baobab/Fig tree	Flowering and generation of new leaves indicate rainfall onset; and prospects of long rains	(Mahoo et al. 2015)
Venonia	When this tree starts to produce flowers, it is an indication of early onset of rains	(Mahoo et al. 2015)
Albizia	The appearance of insects locally known as vidododo on Albizia trees with water dripping from the trees during a dry season is an indication of the coming of short rains	(Mahoo et al. 2015)

Astronomical and meteorological IK indicators for malaria prediction based on rainfall and temperature prediction.

Astronomical and meteorological indicators		
Indicators	Significance	Reference
Moon	When there is a good rain season (short or long rain season) the moon is surrounded by a ring. The bigger the ring, the more rain is expected. A bigger ring signifies a high incidence of malaria in that season. The moon crescent facing upwards indicates upholding water and when facing downwards is releasing water in the next three days. The moon surrounded by heavy clouds indicates a good rainfall season. The slanted position of the moon crescent is an indicator of a bad season, especially towards the slanted side; indicating more diseases and erratic rainfall. The appearance of the moon in a circular shape (normal) is an indication of the onset of rainfall, while the moon with different colours like a rainbow indicates the onset of short and long rains. The red moon indicates an onset of short rains. White moon indicates the onset of rain during the long rainy season. A moon with a halo indicates heavy rainfalls. A dark moon in July accompanied by cold, ice and snowfall signifies a good rainy season.	(Kijazi, Chang'a and Liwenga, 2013; Mahoo et al. 2015; Kolawole et al., 2016; Macherera, Chimbari and Mukaratirwa, 2017; Mafongoya and Ajayi, 2017)
Stars	Star pattern and movement from west to east at night under clear skies means rain will fall in 3 days. Five stars indicate that the rainfall season is approaching	(Mafongoya <i>et al.</i> , 2017) (Kijazi <i>et al.</i> , 2013) (Kijazi, Chang'a and Liwenga, 2013; Mafongoya, and Ajayi, 2017)
Clouds	Clouds that show a good rainy season and high malaria incidence are those thinly spread in the sky. Clouds along the horizon mean much rain that year and a high incidence of malaria.	(Mahoo et al. 2015; Macherera et al. 2017)

	The appearance of many nimbus clouds; and the appearance of red clouds in the morning is an indication of rain in 1 to 3 days. White clouds appearing in the evening indicate the prospect of rain the next day	
Temperatures	When temperatures are very high but alternating with windy conditions, a good rainy season is indicated for that year. High temperatures in October and November signify the onset of rainfall and the prospect of a good rainy season. Malaria incidence is predicted to be high. High temperatures in lowlands in August indicate more rains in the coming season. High temperature in October and November signifies near onset and a good rain season. High temperature during the night is an indication that it will rain the next day	(Kijazi, Chang'a and Liwenga, 2013; Mahoo et al., 2015; Macherera, Chimbari and Mukaratirwa, 2017; Mafongoya and Ajayi, 2017)
Wind	When the wind blows from the east to the west and then from the west to the east there will be much rain in that year and high malaria incidence. If the wind continually blows from the east to the west, there will not be any rain and low malaria incidence. Frequent occurrences of whirlwinds during September and October show that there is going to much rain that year and high malaria incidence. Winds moving from west to east indicate the onset of the rainy season Strong winds in July through October indicate less rainfall in the coming season. The north-south wind direction in September and in late February indicates the onset of short or long rainy season and in most cases a sign of heavy rainfall. Frequent winds indicate that it will rain within one week. Northwest wind direction in November or December Heavy rainfall is going to fall. Also, it indicates due onset of the rainy season. Change in wind direction and temperature Signifies imminent rainfall. Wind blowing without drying vegetation and soil is also an indication of rain. Cold wind for about a week is usually followed by rain.	(Kijazi, Chang'a and Liwenga, 2013; Mahoo et al., 2015; Kolawole et al., 2016; Macherera, Chimbari and Mukaratirwa, 2017; Mafongoya and Ajayi, 2017)
Air and sky	When the air is hazy or misty around the months of September and October, there will be much rain during the coming rainy season and high malaria incidence. When mist covers mountains indicates good rains. The appearance of fog/haze in the morning indicates no rain.	(Soropa <i>et al.</i> , 2015; Macherera, Chimbari and Mukaratirwa, 2017; Mafongoya and Ajayi, 2017)

	Falling of dew early in the morning after the first rains indicates the rainy season.	
Sun	Sun hitting without wind indicating rains	(Mahoo et al. 2015)
Lightning	Repeated lightning in the evening during the dry season is an indication of the onset of short rains	(Mahoo et al. 2015)

IK indicators for malaria prediction based on rainfall and temperature prediction.

Other indicators		
Indicator	Significance	Reference
Human body	The body feels increased or excessive heat during the night and day; a feeling of body pain (headache, flu, back aches) indicates rain in 1-3 days.	(Mafongoya, and Ajayi, 2017)
Water bodies e.g. dams	Maintenance of high water levels in water bodies after the first rains indicates a good rain season ahead.	(Mafongoya and Ajayi, 2017)
Asthmatic attacks	Asthmatic attack, increased pain from past body wounds e.g. from surgical operations means imminent cold and humid conditions.	(Mafongoya, and Ajayi, 2017)
Births	Births of too many girls indicate the rainy season while births of too many boys indicate the dry season.	(Soropa et al. 2015)

Appendix B: Ethical Clearance



FACULTY RESEARCH AND INNOVATION COMMITTEE

RESEARCH ETHICS APPROVAL LETTER

Date: 1 October 2020

This is to confirm that:

Applicant's Name	PM Phoobane
Supervisor Name for Student Project (where applicable)	Prof Masinde
Level of Qualification for Student Project (where applicable)	PhD
Title of research project	Development of an intelligent Prediction Model for Infectious Disease Outbreaks: An Ensemble of Machine Learning, Big Climate Data, and Indigenous Knowledge

Ethical clearance has been provided by the Faculty Research and Innovation Committee on 23 Sep 2020 in view of the CUT Research Ethics and Integrity Framework, 2016 with reference number FEBIT 6.7/6.14- 20-08-2020.

The following special conditions were set:

☒ None

☐ Specific conditions

We wish you success with your research project.



Prof Yali Woyessa
(FRIC Chairperson)

Appendix C: Certificate of Completion of Research Ethics Module.



Zertifikat Certificat

Certificado Certificate

Promouvoir les plus hauts standards éthiques dans la protection des participants à la recherche biomédicale
Promoting the highest ethical standards in the protection of biomedical research participants

Certificat de formation - Training Certificate
Ce document atteste que - this document certifies that

Paulina Maingosinathi Phoobane
a complété avec succès - has successfully completed

Archived - Module 1 - Introduction to Research Ethics
du programme de formation TRREE en évaluation éthique de la recherche
of the TRREE training programme in research ethics evaluation

Release Date: 2020/08/27
CID: PRG/18999a


 Professeur Dominique Sprumont
 Coordinateur TRREE Coordinator

APPROVED BY



Programmes de formation continue (2 crédits)
Continuing Education Programs (2 credits)



Clinical Trials Centre
The University of Hong Kong

Ce programme est soutenu par - This program is supported by :

European and Developing Countries Clinical Trials Partnership (EDCTP) (www.edctp.org) - Swiss National Science Foundation (www.snf.ch) - Canadian Institutes of Health Research (<http://www.cihr-irac.gc.ca/2891.html>) - Swiss Academy of Medical Science (SAMS/ASSM/SAMW) (www.sams.ch) - Commission for Research Partnerships with Developing Countries (www.kfrpc.ch)

[REV : 20240820]

Appendix D: Questionnaire

Development of an intelligent Prediction Model for Infectious Disease Outbreaks: An Ensemble of Machine Learning, Big Climate Data, and Indigenous Knowledge

Part A: Introduction

Local communities use Indigenous Knowledge (IK) for decision making for survival in both the natural and human-built environments (Abera and Science, 2020). Generally, local communities make physical observations and use environmental and astronomical indicators such as the patterns and movement of clouds and objects in the night sky, the flowering or flushing of specific trees and animals' behaviour to understand the environment surrounding them. In this way, they can, therefore, develop mitigation and adaptation strategies to reduce their vulnerability to environmental variabilities, such as climate change and its impact.

The Department of Information Technology at the Central University of Technology, Free State, is conducting research to identify and document the unstructured weather indicators based on the following categories 1) insects, 2) birds, 3) plants, 4) animals, 5) astronomical and meteorological and 6) other indicators.

Phase I of this research seeks to collate the IK from local communities who are IK holders in the **Limpopo province of South Africa**. The results of this research will be used to develop an early warning system that predicts the outbreak of infectious diseases like malaria. You are requested to participate in this research by completing this questionnaire. You are required to put a mark (✓ or X) in the check box for the appropriate option or write down your response in the area provided.

For any questions, please contact the persons whose details are listed at the bottom of the page.

Part B: Consent Form

I, the undersigned, confirm that (please tick box as appropriate):

1. I have read and understood the information about the research. ☐
2. I have been given the opportunity to ask questions about the research and my participation ☐
3. I voluntarily agree to participate in the research. ☐
4. I understand I can withdraw at any time without giving reasons and that I will not be penalized for withdrawing ☐
5. The procedures regarding confidentiality have been clearly explained to me ☐
6. If applicable, separate terms of consent for forms of data collection have been explained and provided to me ☐
7. The use of the data in research, publications, sharing and archiving has been explained to me ☐
8. I understand that other researchers will have access to this data only if they agree to preserve the confidentiality of the data and if they agree to the terms I have specified in this form ☐
9.
 - i) I understand that my responses will be kept strictly confidential ☐
 - ii) I understand that my name will not be linked with the research materials and will not be identified or identifiable in the report or academic publication that result from the research ☐
10. I agree to sign and date this informed consent, along with the Researcher. ☐

Part C: Demographic Information

Q1. Names _____ (optional)

Q3. Gender ☐ Female ☐ Male ☐ Other

Q3. Age

☐ under 18 ☐ 18-35 ☐ 36-45 ☐ 46-55 ☐ 56-65 ☐ above 66

Q4. Highest qualifications

☐ None ☐ Primary ☐ Secondary ☐ Above secondary

Q5. What is the name of your village/town? _____

Q6. What is the main economic activity in your village/town?

Q7. How long have you stayed in this village/town?

☐ 5-10 years ☐ 11-20 ☐ above 20

Q8. Do you own a phone or have access to a phone?

☐ Yes, I own a phone

☐ No, I do not own a phone

☐ I do not own a phone, but I have access to a phone

Q9. If you own a phone, is this a smart phone?

☐ Yes ☐ No

Part D: Knowledge on Weather Forecasting

Q10. Do you check weather forecast information?

☐ Yes ☐ No

If yes, how often do you check it?

☐ Daily ☐ Weekly ☐ Monthly ☐ Seasonal

Q11. Where do you get your weather forecast information? (You may tick more than one box). Do you have confidence in the accuracy of information you get from these options? Please, tick on a scale 1 –5, with one (1) being the lowest level of confidence and five (5) being the highest.

	1	2	3	4	5
Radio	☐	☐	☐	☐	☐
TV	☐	☐	☐	☐	☐
Newspaper/Print media	☐	☐	☐	☐	☐
Website/APP	☐	☐	☐	☐	☐

Local observations.	☐	☐	☐	☐	☐
For example, the clouds and behaviour of Animals or/and plants					
Other. Specify please.	☐	☐	☐	☐	☐

Q12. How long have you been using weather forecast information?

☒ Never ☐ Less than 1 year ☐ 2 to 5 years ☐ Over 5 years

Q13. Does the weather and climate information received adequately meet your requirements? (You may select many requirements or none from the requirements given below). Please, tick on a scale 1 –5, with one (1) being the lowest level of adequacy and five (5) being the highest.

	1	2	3	4	5
Disease prediction (for example, malaria prediction)	☐	☐	☐	☐	☐
Farming	☐	☐	☐	☐	☐
Travelling	☐	☐	☐	☐	☐
Plan the day (like, how to dress, take umbrella or not)	☐	☐	☐	☐	☐
Other. Specify please.	☐	☐	☐	☐	☐

If the weather and climate information received does not adequately meet your requirements, (rating 1 or 2), what is missing?

.....
.....

Q14. Do you receive any weather-related alerts?

☐ yes ☐ No

If yes, please tick source where you MOSTLY receive the weather-related alerts.

	alerts				
Source	Heat stress	Air quality	ultraviolet	malaria	Other. Specify please.....
Radio	☐	☐	☐	☐	☐
TV	☐	☐	☐	☐	☐
Newspaper/printed media	☐	☐	☐	☐	☐

Website/App	☐	☐	☐	☐	☐
Natural indicators (like birds etc.,)	☐	☐	☐	☐	☐
Other. Please indicate	☐	☐	☐	☐	☐

Q15. If weather/climate tailored products, e.g. district level products are generated in a format and at intervals of your choice, would you be willing to pay for these tailored services and products?

☐ Yes ☐ No

Q16. If yes, how much are you willing to pay annually (in Rands)?

☐ R 0-1000

☐ R1001-10 000

☐ R10001-10 0000

Part E: Indigenous Knowledge Indicators for rainfall/Malaria (Focus group)

Q17. Do you use weather forecast to predict malaria occurrence?

Q18. Do you use indigenous knowledge to predict rainfall/drought or malaria occurrence?

Q19. If your answer to Q20 is yes, please list some of the indigenous indicators that you commonly use to make rainfall/drought forecast or malaria prediction.

astronomical and
meteorological; e.g. Visible
phases of full moon signify drier
period

Behaviours of birds		
Behaviours of insects; e.g. ants moving in a straight line indicate a dry spell is imminent		
Behaviours of animals; e.g. cattle coming home jumping with their tails up is a sign of a good season.		
Plants; Flower, leave and Fruit		
Other indicators; Myths and religious beliefs, e.g. an extreme drought is a curse for the offences to the gods		
Knowledge of seasons e.g. it always rains during summer		
Thank you for participating in the questionnaire. Please note that your responses are for research purposes only and will be kept strictly confidential. <i>End of document</i>		